AD_____

Award Number: DAMD17-01-1-0328

TITLE: Computer-aided Characterization of Breast Masses on Volumetric Ultrasound Images: An Adjunct to Mammography

PRINCIPAL INVESTIGATOR: Berkman Sahiner, Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, MI 48109

REPORT DATE: October 2005

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01-10-2005	Annual	6 Sep 2004 – 5 Sep 2005

4. TITLE AND SUBTITLE
Computer-aided Characterization of Breast Masses on Volumetric Ultrasound Images: An Adjunct to Mammography

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-01-1-0328

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Berkman Sahiner, Ph.D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Ann Arbor, MI 48109

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. **ABSTRACT**

The purpose of this project is to develop computer techniques for the analysis of breast ultrasound images, and to combine computerized sonographic and mammographic analyses. The techniques developed in this project are aimed at providing a second opinion to the radiologists in the task of making a biopsy recommendation. In the no-cost time extension year, we have completed and analyzed the observer performance studies proposed for the project. Our results indicate that our CAD system can significantly improve radiologists' likelihood of malignancy estimates. This result has been consistent in the two observer studies that we conducted, the first study involving US volumes only, and the second study involving both US images and mammograms. Our results also indicate that if we select a likelihood of malignancy cutoff threshold for biopsy recommendation with CAD so that the sensitivities with and without CAD are equal, 4 out of 5 radiologists showed significant improvement in their specificity in the first observer study. In the second study, 3 out of 10 radiologists showed significant improvement in their specificity under the same condition. Further improvement of our methods can provide radiologists with a powerful aid for decision making, which may help reduce benign biopsies and improve patient care.

15. SUBJECT TERMS
Computer-aided diagnosis; ultrasonography; breast masses; breast cancer detection

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	71	19b. TELEPHONE NUMBER *(include area code)*

(3) Table of Contents

(4) Introduction

At present, biopsy is the gold standard in breast lesion characterization. However, the positive breast biopsy rate is only about 15-30%. This means that 70-85% of breast biopsies are performed for benign lesions. In order to reduce patient anxiety and morbidity, as well as to decrease health care costs, it is desirable to reduce the number of benign biopsies without missing malignancies. Mammography and sonography are two low-cost imaging modalities that may be improved so that radiologists can obtain more accurate diagnostic information to differentiate malignant and benign lesions. Computerized analysis of the lesions on these images is one of the promising tools that may improve the radiologists' accuracy in characterizing these lesions by providing a consistent and reliable second opinion to radiologists.

In this project, our goal is to analyze volumetric images to improve the accuracy of computerized sonographic breast lesion characterization, and to combine these characterization results with those obtained by computerized analysis of mammograms. Computerized image analysis, feature extraction, and classification methods will be developed to characterize breast masses on three-dimensional or volumetric ultrasound (US) images. The output of the classifier will be a computer rating related to the likelihood of malignancy of the mass. The accuracy of this rating will be studied by comparing it to the biopsy results. We will then combine this rating with a similar rating obtained by computerized analysis of the mammograms of the same patient. The combined classifier is expected to be more accurate than either classifier alone.

(5) Body

In the current project year (9/6/04-9/5/05), we have performed the following studies:

(A) Observer experiment with CAD on 3D US images

Earlier in our project, we conducted a study to investigate if our computer classifier that uses 3D US volumes would improve radiologists' accuracy in differentiation of malignant and benign breast masses on ultrasound images. The results of this analysis were also submitted to the journal Radiology as an original research paper [1] (Appendix 1). This paper is still under review after revisions performed in the current year. We present below the additional analyses performed in the no-cost time extension (NCTX) period.

The data set, methods and the results for this study are explained in detail in Appendix 1. Our previously developed computer algorithm [2] had an area under the Receiver Operating Characteristic (ROC) curve of A_z=0.92. Five radiologists (RAD1-RAD5) participated in this study as observers. They read the 3D US images using a specially-developed software first without CAD and then with CAD. They provided a likelihood of malignancy (LM) rating under both conditions. The LM ratings of the radiologists with and without CAD were analyzed using ROC methodology [3].

In the NTCX period, we analyzed the data to investigate how heavily the radiologists rely on the computer results. When an observer experiment is performed to investigate the impact of CAD on radiologists' decisions, there may be a concern that, in a laboratory environment, the radiologists may rely too heavily on the CAD system without adequately merging the computer output with their own judgment. To investigate whether this is the case, we estimated the correlation between the radiologists' readings with CAD to (i) their readings without CAD, and (ii) the computer scores. We then estimated the statistical significance of the difference between these two correlation coefficients using Cohen and Cohen's method [4]. If the radiologists utilize the computer scores only when they believe it makes a true contribution to their original assessment, then the correlation (i) above should be significantly higher than (ii).

The correlations between the radiologists' readings with and without CAD, and those between radiologists' readings with CAD and computer scores are shown in Table 1. The former of these correlations was higher than the latter for all five radiologists, and the difference between the two was statistically significant ($p<10^{-6}$) for four radiologists. This result indicates that when they read with CAD, radiologists had a higher agreement with their reading before CAD compared to their agreement with computer scores.

Table 1.

Correlations between the radiologists' LM ratings with and without CAD, and those between radiologists' LM ratings with CAD and computer scores

Rad. No	Correlation between LM ratings with CAD and		p-value
	LM ratings w/o CAD	Computer scores	
1	0.94	0.70	$<10^{-6}$
2	0.96	0.61	$<10^{-6}$
3	0.96	0.72	$<10^{-6}$
4	0.86	0.83	0.26
5	0.94	0.70	$<10^{-6}$

We also investigated whether further combining the computer scores with the radiologists' readings with CAD may improve the accuracy in the characterization task. If the radiologists excessively rely on the computer scores, then such a combination should not improve the

accuracy beyond that of the computer classifier. The computer scores were combined with the radiologists' LM ratings with CAD by first normalizing the computer scores to the same range as the radiologist's scores (0-100) and then averaging. The average scores were analyzed using ROC methodology. The resulting A_z values are listed in Table 2. It is observed that by combining the scores in this manner, the A_z values of four out of five radiologists can be improved beyond that of the computer classifier alone. Using the DBM method, it was also found that the average A_z value over all radiologists (=0.94), if this strategy was used, would be significantly higher (p=0.008) than the average A_z (=0.90) from the radiologists' LM ratings with CAD obtained in our observer experiment.

Table 2.

The A_z values when the radiologist LM ratings with CAD are averaged with computer scores

R1	R2	R3	R4	R5	Computer score
0.94	0.92	0.94	0.95	0.94	0.92

When radiologists perform a laboratory experiment with CAD, there may be a concern that they may overly rely on the computer results, especially if they know that the computer classifier is very accurate. However, our observations indicated that the radiologists did not develop a trend to follow the computer ratings in this study. First, we discussed in our previous yearly report, the radiologists did not change their LM rating substantially (i.e., greater than or equal to 5 on the 100 point scale) with CAD in 64% (326/510) of the readings. Second, the correlation analysis (Table 1) indicates that the LM ratings of a radiologist with and without

CAD were highly correlated, whereas the correlation between "computer scores" and their "LM ratings with CAD" was significantly lower for four readers. Third, when the radiologist LM ratings with CAD were further combined with the computer scores by averaging, the A_z values of four of the radiologists were higher than that of the computer classifier. Had the radiologists excessively relied on the computer, one would not expect these combined scores to be more accurate than the computer classifier. We therefore conclude that the radiologists did not rely on the computer scores excessively.

In the NCTX period, we also statistically analyzed the change in radiologists' sensitivity and specificity with CAD. For this purpose, we employed McNemar's test [5] by considering the number of beneficial and detrimental changes with CAD in biopsy recommendation for the malignant cases. If a malignant case was not recommended for biopsy without CAD, but was recommended for biopsy with CAD, this was defined as a beneficial change. If a malignant case was recommended for biopsy without CAD, but was not recommended for biopsy with CAD, this was defined as a detrimental change. We similarly applied McNemar's test to benign cases to investigate whether the change in specificity with CAD is statistically significant. McNemar's test was applied to under two kinds of conditions: In the first condition, we used a 2% LM threshold for reading both without and with CAD. Under this condition, on the average, we observed an improvement in the sensitivity (0.96 and 0.98 without and with CAD respectively) and a decrease in specificity (0.22 and 0.19 without and with CAD, respectively). Under condition #1, the change in the sensitivity and specificity did not achieve statistical significance for any radiologist (range of p-values using McNemar's test: 0.157-1.00 for sensitivity, and 0.102-1.00 for specificity). Under the second condition, we adjusted the LM threshold for reading with CAD to 7%. At this threshold, the average sensitivity both without

and with CAD were 96%, so that the comparison for specificity with and without CAD are performed at corresponding operating points on the ROC curve. Under this condition, the improvement in specificity for 4 out of 5 radiologists was statistically significant ($p<0.002$, McNemar's test), while the change in sensitivity for each radiologist was insignificant.

(B) Observer experiment with CAD on 3D US images and mammograms

In the NCTX period, we completed our observer study for evaluating the effect of the multimodality computer classier on radiologists' accuracy for the characterization of masses on US volumes and mammograms. The details of the data set and the computer classification method is provided in out previous yearly report. Briefly, we had 32 benign and 35 malignant masses in our data set. The total number of mammographic views was 163, with each case containing between one and three views (CC, MLO, or LAT). Ten radiologists read the cases sequentially under three conditions. First, the radiologist read the mammogram regions of interest (ROIs), and provided a BI-RADS (Breast Imaging Reporting and Data System) score and a likelihood of malignancy rating. Second, the US images were displayed along with the mammogram ROIs, the radiologist provided a second malignancy rating, and recommended: (i) 1-year follow-up; (ii) short-term follow-up; or (iii) biopsy. Third, the computer score was displayed, and the radiologist provided a third malignancy rating and revised the recommended action. The radiologist ratings were analyzed using ROC analysis. We also analyzed the sensitivity and specificity under the three different reading conditions.

Figure 1 shows the A_z values of each radiologist under the three reading conditions. It is observed that for each radiologist, condition 3 (reading with CAD) was the most accurate, followed by condition 2 (reading the US and mammograms images without CAD) and then followed by condition 1 (reading mammogram images alone). The A_z value of the computer

(0.92) is also shown as a dotted line. It is observed that six of the radiologists' A_z values under condition 2 are higher than that of the computer classifier's. Despite this, all of the radiologists showed improvement when they read with CAD.

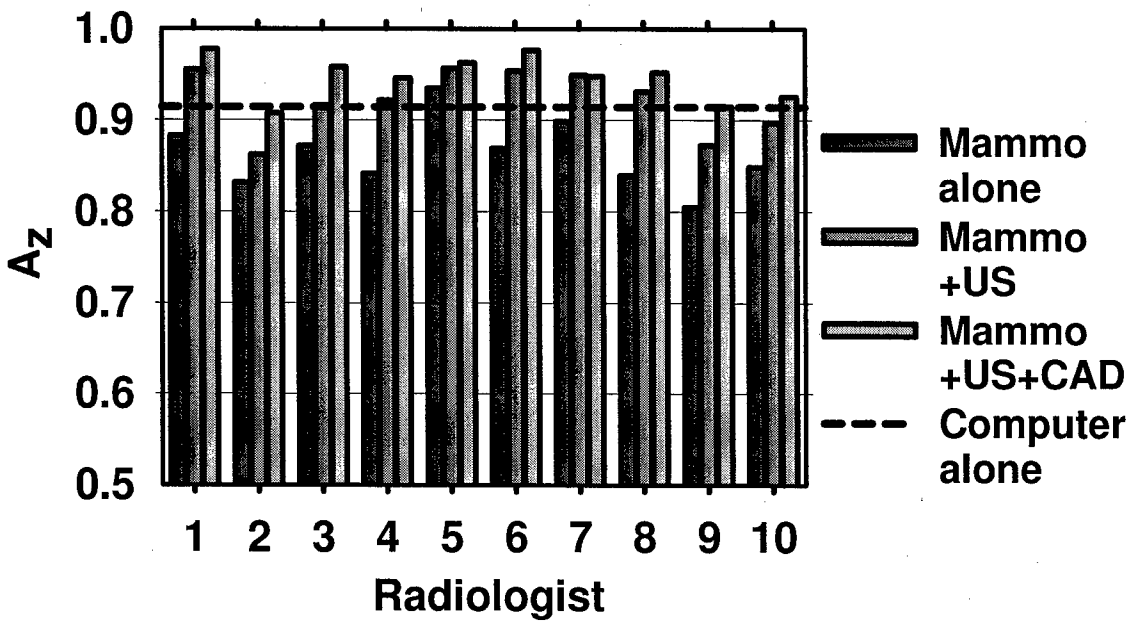


Figure 1: The area A_z under the ROC curve for each radiologist under the three reading conditions.

Figure 2 shows the average ROC curves obtained by averaging the a and b values under each of the reading conditions for the ten radiologists. The average A_z values under the conditions (1), (2), and (3) were 0.87, 0.93, and 0.95, respectively.

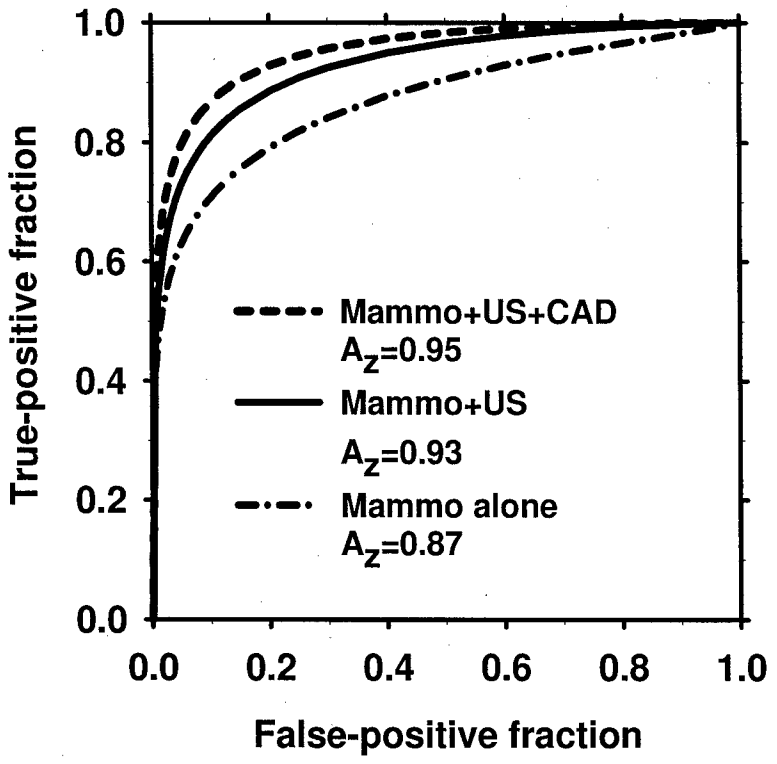


Figure 2: The average ROC curves for the ten radiologists under the three reading conditions.

To investigate whether the improvement with CAD is statistically significant, we used Dorfman Berbaum-Metz multiple reader multiple case (MRMC) analysis [6], as well as the t-test. Both methods indicated that the improvement is statistically significant (p=0.05 using MRMC method and p=0.0005 using the t-test).

The sensitivities and specificities of the ten radiologists are shown in Table 3. The average sensitivity with CAD improved from 0.98 to 0.99, while the average specificity improved from 0.27 to 0.29. An important difference between this study and the previous study that used the US images only is that we observed an improvement in both sensitivity and specificity with CAD when both US volumes and mammography were included in the study, although both improvements were small. In the previous study (Section A), we had found that

when only mammograms are included, the use of CAD improved the sensitivity, but decreased the specificity of the radiologists.

Table 3: The sensitivity and specificity of each radiologist under the three reading conditions.

	Mammography alone		Mammography +US		Mammography +US+CAD	
Rad. #	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
01	0.97	0.34	1.00	0.16	1.00	0.16
02	0.83	0.66	0.89	0.66	0.94	0.63
03	1.00	0.13	1.00	0.13	1.00	0.19
04	0.91	0.41	0.97	0.34	1.00	0.34
05	1.00	0.16	1.00	0.16	1.00	0.28
06	0.91	0.44	0.97	0.44	1.00	0.34
07	1.00	0.13	1.00	0.09	0.97	0.28
08	0.77	0.84	0.94	0.53	0.97	0.50
09	1.00	0.00	1.00	0.09	1.00	0.09
10	1.00	0.25	1.00	0.13	1.00	0.09
Avg.	0.94	0.33	0.98	0.27	0.99	0.29

To statistically analyze the change in radiologists' sensitivity and specificity with CAD, we employed McNemar's test. The test was applied under two kinds of settings: In the first setting, we used the BI-RADS descriptors and the action categories provided by the radiologists under conditions (2) and (3). In this setting, none of the differences in sensitivities of specificities with and without CAD was statistically significant for any of the radiologists. Under the second setting, we adjusted the LM threshold for reading with CAD to 9%. At this threshold, the average sensitivity both without and with CAD were 98%, so that the comparison for specificity with and without CAD are performed at corresponding operating points on the

ROC curve. Under this condition, the improvement in specificity for 3 out of 10 radiologists was statistically significant.

Our observer studies therefore indicated some of the advantages and disadvantages of our CAD method for characterization of breast masses. Both studies have shown that the radiologists' ROC curve improve significantly when they use CAD. The improvement especially in the second study is remarkable because the radiologists were already very accurate when they combined the information on the US images and mammograms without CAD. Although the computer was less accurate (A_z=0.92) than the average radiologist (A_z=0.93) the radiologists were still able to merge the computer information with their own assessment in a significantly beneficial way. However, the results for sensitivity and specificity were less remarkable. Although we observed an improvement in sensitivity with CAD in both studies, the improvement did not reach statistically significance when we used a 2% LM threshold for biopsy recommendation. This is the threshold currently recommended by the American College of Radiology (ACR). We also compared specificities under the hypothetical condition that the LM threshold is chosen such that the sensitivities with and without CAD are the same. Under this condition, we had significant improvement in specificities for some, but not all of the radiologists. This result underlines that to be clinically useful, we may have to further increase the accuracy of our computer classifier so that radiologists have more confidence in our system and the computer can provide additional complementary information to the radiologist.

The development of computer-aided diagnosis algorithms is an iterative process. It generally takes many iterations in order to bring the performance of the automated methods up to the acceptable level and to work reliably. In the process of designing the specific computer classifier algorithms in this project (pre-processing, segmentation, feature extraction,

classification, etc.), we have conceived many ideas about how the performance of the computer classifier can be improved. For example, in the past year, through other projects, we have been gaining experience in a new segmentation technique called "level-set methods", and we believe that applying this technique to the ultrasound data will provide improved results. We have also identified several feature extraction techniques that hold great promise in our project. Based on the experience we have gained in the past years of our project, we believe that performing another iteration to integrate these new techniques into our CAD system will further improve its accuracy.

(6) Key Research Accomplishments

- We performed further analysis of the observer performance study conducted earlier (radiologists reading 3D US volumes without and with CAD). Our analysis showed that although the accuracy of the computer classifier for 3D ultrasound images was higher than that of the radiologists, the radiologists did not excessively rely on the computer scores
- The analysis of sensitivity and specificity for the observer study involving mammograms alone indicated that reading with CAD did not result in significantly higher sensitivity or specificity if a clinically accepted likelihood of malignancy threshold of 2% is used for biopsy recommendation. However, when the LM threshold for reading with CAD is adjusted to 7% so that the average sensitivity both without and with CAD were 96%, four out of five radiologists showed significant improvement with CAD in their specificity.
- We completed the observer performance study with multi-modality CAD (US volumes and mammograms). MRMC analysis indicated that the accuracy of LM ratings of radiologists was significantly improved when they read the images with CAD.

- The analysis of sensitivity and specificity for the observer study involving multi-modality imaging indicated that reading with CAD did not result in significantly higher sensitivity or specificity if a clinically accepted likelihood of malignancy threshold of 2% is used for biopsy recommendation. However, when the LM threshold for reading with CAD is adjusted to 9% so that the average sensitivity both without and with CAD were 98%, three out of ten radiologists showed significant improvement with CAD in their specificity.

(7) Reportable Outcomes

The journal paper submitted to Radiology on the effect of the 3D US classifier on radiologists' characterization of breast masses on ultrasound images has been revised and resubmitted. Additionally, we presented our results at RSNA 2004. We are in the process of writing a manuscript for journal submission based on this conference abstract.

Journal Publications:

Sahiner B, Chan HP, Roubidoux MA, Hadjiiski L, Helvie MA, Paramagul C, Bailey J, Nees A, Blane C, "Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy," *Radiology (resubmitted)* 2005.

Conference Abstracts:

Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul CP, Helvie MA, et al., "The effect of a multi-modality computer classifier on radiologists' accuracy in characterizing breast masses using mammograms and volumetric ultrasound images: An ROC study," presented at the *90th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, Nov. 28-Dec 3, 2004.

Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul CP, Helvie MA, "Computer-aided multi-modality breast mass characterization," presented at the fourth Era of Hope meeting for the Department of Defense (DOD) Breast Cancer Research Program (BCRP), June 8-11, 2005, Philadelphia, Pennsylvania.

(8) Conclusions

In the NCTX year of the USAMRMC BCRP grant, we have completed and analyzed the observer performance studies proposed for the grant. Our results indicate that our CAD system can significantly improve radiologists' likelihood of malignancy estimates. This result has been consistent in the two observer studies that we conducted, one involving US volumes only, and the other involving both US images and mammograms. To quantify the improvement, we used the ROC methodology and the MRMC method. For more immediate clinical impact of CAD however, the sensitivity and specificity of the radiologists with and without CAD are more important than the likelihood of malignancy estimates. Our results indicate that if we select a likelihood of malignancy cutoff threshold for biopsy recommendation with CAD so that the sensitivities with and without CAD are equal, then 4 out of 5 radiologists showed significant improvement in their specificity in the observer study involving US volumes only. In the study involving both US images and mammograms, 3 out of 10 radiologists showed significant improvement in their specificity under the same condition. When the currently accepted 2% LM cutoff, we observed a trend for improvement, but the change did not reach statistical significance for any of the radiologists. Further improvement of the 3D ultrasound characterization methods and improved methods for combination with mammographic computer image analyses can provide radiologists with a powerful aid for decision making, which may help reduce unnecessary biopsies and improve patient care.

(9) References

1. B. Sahiner, H.P. Chan, M.A. Roubidoux, L.M. Hadjiiski, M.A. Helvie, C. Paramagul, J. Bailey, A. Nees, and C. Blane, "Computer-aided diagnosis of malignant and benign breast masses in 3d ultrasound volumes: effect on radiologists' characterization accuracy," Radiology, (Submitted), (2005).
2. B. Sahiner, H.P. Chan, M.A. Roubidoux, M.A. Helvie, L.M. Hadjiiski, A. Ramachandran, G.L. LeCarpentier, A. Nees, C. Paramagul, and C.E. Blane, "Computerized characterization of breast masses on 3-D ultrasound volumes," Medical Physics, 31, 744-754 (2004).
3. J.A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," Invest. Radiology, 14, 109-121 (1979).
4. J. Cohen and P. Cohen, *Applied multiple regression/correlation analysis for the behavioral sciences*. 1983, Hillside, NJ: Lawrence Erlbaum.
5. Q. McNemar, "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages," Psychometrika, 12, 153 -157 (1947).
6. D.D. Dorfman, K.S. Berbaum, and C.E. Metz, "ROC rating analysis: Generalization to the population of readers and cases with the jackknife method," Investigative Radiology, 27, 723-731 (1992).

(10) Appendix

Copies of the following publications are enclosed with this report:

(1) Sahiner B, Chan HP, Roubidoux MA, Hadjiiski L, Helvie MA, Paramagul C, Bailey J, Nees A, Blane C, "Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy," *Radiology (resubmitted)* 2005.

(2) Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul CP, Helvie MA, et al., "The effect of a multi-modality computer classifier on radiologists' accuracy in characterizing breast masses using mammograms and volumetric ultrasound images: An ROC study,"

presentated at the *90th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, Nov. 28-Dec 3, 2004.

(3) Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul CP, Helvie MA, "Computer-aided multi-modality breast mass characterization," presented at the fourth Era of Hope meeting for the Department of Defense (DOD) Breast Cancer Research Program (BCRP), June 8-11, 2005, Philadelphia, Pennsylvania.

Appendix 1

Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy

Authors:

Berkman Sahiner, Ph.D.
Heang-Ping Chan, Ph.D.
Marilyn A. Roubidoux, M.D.,
Lubomir M. Hadjiiski, Ph.D.
Mark A. Helvie, M.D.
Chintana Paramagul, M.D.
Janet Bailey, M.D.
Alexis Nees, M.D.
Caroline Blane, M.D.

Institutional Affiliations:

Department of Radiology
The University of Michigan
CGC B2102,
1500 East Medical Center Drive
Ann Arbor, MI 48109-0904

Corresponding Author:

Berkman Sahiner, Ph.D.
Phone: 734-647-7429
Fax: 734-615-5513
Email: berki@umich.edu

Grant Information:

This work was supported by USAMRMC grant DAMD17-01-1-0328 and by USPHS grants CA095153 and CA91713.

RSNA Presentation:

This study was partly presented at the RSNA 2003 meeting (Abstract #504)

Type of manuscript:

Original research

Computer-Aided Diagnosis of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: Effect on Radiologists' Characterization Accuracy

Type of manuscript:
Original research

Advances in Knowledge:

The potential improvement in radiologists' characterization accuracy of breast masses in 3D ultrasound volumes was evaluated when they were assisted with an automated computer classifier.

Our results indicate that the CAD algorithm used in this study was able to assist even expert breast imaging radiologists in characterizing masses on 3D US volumes. The average A_z value improved significantly ($p=0.005$) from 0.84 to 0.90, and the average $A_z^{(0.9)}$ value improved significantly ($p=0.015$) from 0.30 to 0.47.

Our data set contained 102 lesions, of which 96 were categorized as solid by the majority rule. When we confined our ROC analysis to the subset of solid masses, the results were virtually unchanged from the entire set of 102 masses.

With CAD, the average likelihood of malignancy (LM) rating decreased for benign masses, and increased for malignant masses. For benign masses, the average decrease in LM rating was 0.77, which did not achieve statistical significance (two-tailed $p=0.51$). The increase in the average LM rating of malignant masses was 5.59, which was statistically significant (two-tailed $p<0.0001$).

ABSTRACT

Purpose: We previously developed an automated computer classifier to characterize breast masses in 3D ultrasound volumes. The purpose of this study was to investigate whether computer aided diagnosis (CAD) using this classifier would improve radiologists' accuracy.

Materials and Methods: Informed consent and institutional review board approval were obtained. Our data set contained 3D ultrasound volumes from 102 cases of biopsy-proven breast masses (46 benign and 56 malignant). A computer algorithm was designed to automatically delineate the mass boundaries and extract features based on the segmented mass shapes and margins. The features were merged into a malignancy score using a computer classifier. Five experienced radiologists participated as readers. Each radiologist read the cases first without CAD, immediately followed by reading with CAD. The observers' malignancy rating data were analyzed using receiver operating characteristic (ROC) methodology.

Results: Without CAD, the five radiologists had an average area under the ROC curve, A_z, of 0.84 (range: 0.81 to 0.87). With CAD, their average A_z increased significantly ($p=0.006$) to 0.90 (range: 0.86 to 0.93). Using a 2% likelihood of malignancy as the threshold for biopsy recommendation, the average sensitivity of the radiologists increased from 96% to 98% with CAD, while their average specificity for this data set decreased from 22% to 19%. If a biopsy recommendation threshold could be chosen

such that the sensitivity were maintained at 96%, the specificity would increase to 46% with CAD.

Conclusion: A well-trained computer algorithm may improve radiologists' accuracy in distinguishing malignant from benign breast masses in 3D ultrasound volumes.

Key Words: Computer-Aided Diagnosis, ROC Observer Study, Classification, Ultrasound, Malignancy.

INTRODUCTION

In current clinical practice, the positive biopsy rate for breast cancer is about 15-30% (1-3). To reduce patient anxiety and morbidity, as well as to decrease health care costs, it is desirable to reduce the number of benign biopsies without missing malignancies. Previous studies on mammography have shown that radiologists' accuracy in distinguishing malignant from benign masses can significantly improve when they use a well-trained computer-aided diagnosis (CAD) system as a second opinion (4-6).

Ultrasound (US) is an important imaging modality for characterization of breast masses. For the differentiation of simple cysts from other lesions, interpretation of US images by experienced breast radiologists results in an accuracy close to 100% (7). In current clinical practice, if a palpable or mammographically suspicious mass cannot be confidently categorized as a cyst in US examination, it is often recommended for biopsy. Several recent studies (8-10) have indicated that the improvement in US imaging technology and the expert interpretation by radiologists may make it possible to characterize solid breast masses as malignant and benign with high accuracy. In a recent publication, Taylor et al. (10) reported that the addition of US evaluation to mammography alone increased the specificity in their data set of 761 biopsy-proven masses from 51.4% to 63.8%, while slightly increasing the sensitivity from 97.1% to 97.9%.

Several groups of researchers have been developing methods for computerized characterization of masses on 2-dimensional US images (11-14). We have recently developed an automated computer classifier for differentiation of malignant and benign

breast masses in 3-dimensional (3D) US volumes (15). The purpose of this study was to investigate the effect of our computer classifier on radiologists' accuracy in discriminating between malignant and benign masses using 3D volumetric ultrasound images. Both the radiologists and the CAD algorithm analyzed 3D volumetric images of the masses which had been saved as cine-loops. To our knowledge, this is the first observer study to evaluate the impact of a CAD algorithm designed for 3D US images on radiologists' accuracy.

MATERIALS AND METHODS

Data Set

The data collection protocol was approved by our Institutional Review Board prior to the commencement of the study. Individual patient informed consent was obtained from all subjects. The group consisted of 130 consecutive patients who agreed to have a 3D breast US examination between 1998 and 2002. All patients had a sonographic mass assessed as suspicious or highly suggestive of malignancy and were scheduled for biopsy or needle aspiration. Twenty-eight patients from this study group were excluded from our analysis for the following reasons: those who had prior biopsy in the same region of the breast, those with sonographically simple cysts, those with scans which were deemed technically unsuccessful because of motion or other artifacts, and masses which were incompletely imaged in any dimension because of large size or eccentric position in the scan. Thus our study group consisted of 102 patients (average age: 51 years, range: 25-86). Based on biopsy or fine needle aspiration results, 56 masses were malignant and 45 were benign. One of the masses resolved after imaging, and the patient was cancer-free after three year follow-up. Forty-three of the malignancies were

invasive ductal carcinoma, 5 were invasive lobular carcinoma, 3 were ductal carcinoma in-situ, one was medullary carcinoma, and 4 were other invasive carcinoma. Of the biopsy-proven benign masses, 18 were fibroadenoma, 12 were fibrocystic disease, 8 were cyst, 2 were fat necrosis, 2 were scar tissue, one was fibrosis, one was granuloma, and one was other benign breast tissue. The mean mass diameter was 1.28 cm (standard deviation = 0.78 cm).

The 3D US data were acquired using an experimental system that was previously developed and tested at our institution (16, 17). The 3D system consisted of a commercially available GE Logiq 700 (Milwaukee, WI) US scanner with an M12 linear array transducer, a mechanical transducer guiding system, and a computer workstation. The linear array transducer was operated at 11 MHz. The technologist was free to set the focal distance and the overall gain adjustment to obtain the best possible image. Before 3D image acquisition, the technologist used clinical US and mammogram images and reports to identify the suspicious mass. During 3D image acquisition, the technologist manually translated the transducer linearly in the cross-plane, or the z-direction, while the image acquisition system recorded 2D B-mode images in the image scan plane (x-y plane). The 2D images were obtained at approximately 0.5 mm incremental translations, which were measured and recorded using a translation sensor. The scanned breast region measured typically 4.5 cm long by 4.0 cm wide by 4.0 cm deep. The typical in-slice pixel size was approximately 0.11 mm X 0.11 mm.

The B-mode images were recorded into a buffer in the US scanner. After data acquisition, the images and the position data were transferred digitally to a workstation, where individual planes were cropped and stacked to form a 3D volume. The biopsy-

proven mass in each volume was identified by an MQSA (Mammography Quality Standards Act) qualified radiologist (MAR), referred to as RAD0 in the following, using clinical US and mammographic images to confirm that the 3D images contained the mass of interest and showed the mass in its entirety.

Computerized Classification of Masses in US Volumes

The details of our CAD system developed for the classification of masses in 3D US volumes can be found in the literature (15). A summary of the method is provided below.

The first step of the CAD system involved the extraction of the mass boundaries in the 3D volume, i.e., mass segmentation. Automated segmentation of breast masses on US images is a difficult task because of image speckles, posterior shadowing, and the variations of the gray level both within the mass and in the normal breast tissue. We developed a 3D active contour model for segmentation. The active contour model combined the prior knowledge about the relative smoothness of the 3D mass shape in US volume with the information in the image data. An example of the segmented mass slices for a malignant mass is shown in Figure 1.

After mass segmentation, image features were extracted from the mass and its margins for classification. Our feature space consisted of width-to-height ratio, posterior shadowing, and texture descriptors. The mass shape in terms of relative width to height was described by the ratio of the widest cross section of the automatically segmented lesion shape to the tallest cross section. Posterior shadowing features were defined in terms of the normalized average gray-level values in strips posterior to the mass.

Texture features were extracted from two disk-shaped regions containing the boundary of each mass, as well as presumably mass and normal tissue adjacent to the boundary of the mass. These regions followed the contour determined by the active contour model. An illustration of the regions used for computing the posterior shadowing and texture features is shown in Figure 2. Additional details about the feature definitions can be found in the Appendix.

The features described above were extracted from each slice of the US volume containing a mass to define slice-based features. For a given mass, features extracted from different slices were combined to define case-based features. Linear discriminant analysis (LDA) with stepwise feature selection (18) was applied to the case-based feature vectors to obtain computer-estimated malignancy scores. A leave-one-case-out resampling method (19) was used for training and testing of the classification system. The test scores obtained by the leave-one-out partitioning method were used as the malignancy scores in the observer performance study. Two Gaussian functions were fitted to the distributions of the malignancy scores of the benign and malignant classes separately, and were used in the observer performance study as described below.

Observer Performance Study

Five radiologists (MAH, CP, JB, AN, CB), different from the one who was involved in data set collection (RAD0), participated as observers. These five radiologists, referred to as RAD1-RAD5 in the following, had an average of 13 years of experience in mammographic and breast US interpretation (range: 3-26 years) in an academic radiology department at a National Cancer Institute-designated comprehensive

cancer center. They were all MQSA qualified. Four were fellowship-trained in breast imaging, and one had 26 years of experience in breast imaging. At our department, about 4300 breast US examinations are performed annually.

An interactive graphical user interface (GUI), shown in Figure 3, was developed to facilitate the navigation through the scanned 3D volumes of interest containing the mass, and to adjust the window and level of the displayed images. The location of the mass of interest, as determined by RAD0 using all available imaging and pathological information, was marked on each slice, so that all the radiologists would rank the same mass and ignore others if more than one mass could be seen in the volume.

During the experiment, an observer first read a case without CAD. This involved assessing mass characteristics in six categories such as shape, margins, echogenicity, cystic versus solid appearance, and through transmission using the GUI, and providing an estimate of the likelihood of malignancy (LM) for the case on a scale of 0 to 100%. A button corresponding to an LM rating of 0% was provided for benign masses, and another button corresponding to LM ratings of less than 2% was provided for probably benign masses. This second button was set to correspond to the ACR-BIRADS category 3 (probably benign finding) for which short-interval follow-up is recommended (20). The radiologists used a slide bar to enter their ratings between 3% and 100%. The discrete buttons facilitate the selection of these LM ratings more precisely for the benign and probably benign masses because our previous experiences indicate that the uncertainty of selecting ratings on a slide bar by observers can be much greater than 2%. The observers were reminded at the beginning of the study that if they rated a mass as

having larger than 2% of LM, it would indicate that they would recommend the mass for biopsy (20, 21).

We used a two-step sequential reading design, which was found to be a sensitive probe of the difference in the two conditions in previous studies (6, 22). The radiologist first read the US volume without CAD, and rendered an estimate of the LM. The estimate without CAD was stored in a computer file, and the radiologist was not able to modify it after seeing the computer results. Immediately after reading without CAD, the computer-estimated malignancy score for the case was displayed on the screen, and the radiologist rendered an estimate of the LM with CAD. The computer's malignancy score is on a relative rating scale and cannot be converted to the likelihood of malignancy of the masses without making assumptions on the disease prevalence and that the data set at hand is statistically similar to the patient population. To avoid making assumptions, we linearly mapped and rounded the computer's malignancy score to an integer between 1 and 10 before displaying the score on the GUI. In order to provide a reference of the computer performance to the radiologists, the fitted Gaussian distributions to the computer scores for the malignant and benign classes were also displayed on the interface. The radiologists had the option to keep their original malignancy rating, or change it using the slide bar after taking into consideration the computer's opinion. The radiologists were not informed about whether a case was malignant or benign during or after the study, and the overall results of their assessment were not discussed with the radiologists before the study was completed.

There was no time limit for the radiologists. The radiologists were told that practically all of the cases in the study had undergone biopsy, but were not informed

about the proportion of malignant cases in the data set. The case reading order was randomized for each radiologist. In order to reduce the effect of fatigue on the radiologists' performance, the data set was read in three separate sessions by each radiologist. Before participating in the study, the radiologists were trained on five cases that were not part of the test data set. They were familiarized with the study design, the functions on the GUI, and the computer's relative malignancy rating scale during the training session.

The data set used in this investigation was also used in an earlier study for the development of the CAD technique (15). Three of the radiologists (R1, R2, and R3) in the current investigation had already provided an LM rating for these cases without CAD in our earlier study (identified as R3, R4, and R2 in our earlier study) that had a different experimental design and using a different GUI. The two readings by the same radiologists were separated by more than six months. The radiologists were not informed about whether a case was malignant or benign during or after the previous study. The accuracies of these radiologists without CAD in these two studies were compared.

Data Analysis

There is no ground truth for the mass characteristics such as echogenicity and through transmission, since they are judged subjectively by radiologists. To summarize the assessments of the mass characteristics, a "majority assessment" for each category was determined according to the majority rule by the six radiologists (RAD0-RAD5). The majority rule determined which one of the descriptors was selected by the largest number of radiologists. For example, if one radiologist described the echogenicity

characteristics of a mass as hypoechoic, three as markedly hypoechoic, one as anechoic, and one as heterogeneous, the majority assessment for echogenicity of the mass would be markedly hypoechoic. When there was a tie between two descriptors, we used the descriptor chosen by RAD0, who was very familiar with the cases due to the role in data collection, as the tie-breaker. If there was a tie, and the original descriptor provided by RAD0 was not one of the descriptors that were tied, RAD0 was asked to re-read the images and choose one of the tied descriptors.

The LM ratings of the radiologists with and without CAD were analyzed using ROC methodology (23, 24). The area under the ROC curve, A_z, and the partial area index above a sensitivity of 0.9, $A_z^{(0.9)}$ (25) were used as the accuracy measures. For an individual radiologist, the significance of the change in accuracy with CAD was also analyzed using ROC methodology. For the group of five radiologists, the significance of the change in accuracy with CAD was tested using the Dorfman-Berbaum-Metz (DBM) multi-reader multi-case (MRMC) methodology (26) and also using Student's two-tailed paired t-test. The DBM method is normally the preferred method to analyze the A_z values for MRMC data because it accounts for both reader and case variances, while the t-test does not account for case variance in its calculation of the p value. Therefore the conclusions from the t-test are generalizable to the population of readers, but not to the population of cases. The t-test was applied to the evaluation of the partial A_z index above a sensitivity of 0.9. For this task, there is no available software that accounts for both reader and case variances.

The sensitivity and specificity of each radiologist with and without CAD were compared using an LM rating of 2% as the threshold above which biopsy would be

recommended (20, 21). The radiologists in our study are familiar with BI-RADS recommendations and are well aware that selecting an LM>2% is equivalent to declaring that the mass is suspicious enough to warrant biopsy. If the radiologist intended to indicate an LM less than 2%, then he/she would select one of two GUI buttons designated as "benign" and "less than 2% likelihood of malignancy", with the buttons clearly labeled as "benign" and "probably benign". The use of the BI-RADS lexicon and the clear definition of the buttons therefore would record the radiologists' assessment unambiguously, as opposed to a question in text form to the radiologist whether he/she would recommend the case for biopsy without a direct reference to the LM.

In addition to an LM rating of 2%, we also tested a hypothetical biopsy threshold of LM with CAD. This hypothetical threshold was chosen to maintain the average sensitivity of the radiologists at the same level as that without CAD. We could then evaluate the change in specificity if the sensitivity was kept the same before and after use of CAD.

To investigate whether the change in sensitivity with CAD is statistically significant for a given radiologist, we employed McNemar's test by considering the number of beneficial and detrimental changes with CAD in biopsy recommendation for the malignant cases. If a malignant case was not recommended for biopsy without CAD, but was recommended for biopsy with CAD, this was defined as a beneficial change. If a malignant case was recommended for biopsy without CAD, but was not recommended for biopsy with CAD, this was defined as a detrimental change. We similarly applied McNemar's test to benign cases to investigate whether the change in specificity with CAD is statistically significant.

In addition to analyzing the change with CAD in the number of cases for which the LM rating moved across the biopsy threshold of 2%, we also examined the number of cases for which CAD resulted in a substantial change in the LM rating. We defined a substantial change as an absolute value difference of larger than or equal to 5 between LM ratings with and without CAD. The substantial decreases and increases in the ratings of malignant and benign cases were examined. For each mass, we also averaged the changes in the LM ratings by the five radiologists, and compared how CAD changes the average LM ratings for malignant and benign masses.

When an observer experiment is performed to investigate the impact of CAD on radiologists' decisions in a laboratory environment, there may be a concern that the radiologists may rely too heavily on the CAD system without adequately merging the computer output with their own judgment. To investigate whether this is the case, we estimated the correlation between the radiologists' readings with CAD and (i) their readings without CAD, and (ii) the computer scores. We then estimated the statistical significance of the difference between these two correlation coefficients using Cohen and Cohen's method (27). If the radiologists utilize the computer scores only when they believe that it makes a true contribution to their original assessment, then the correlation (i) above should be significantly higher than (ii).

RESULTS

A total of 96 masses were categorized as solid according to the majority rule. Five masses were categorized as complex cysts, and one as a simple cyst by three or more radiologists. One mass that was categorized as a complex cyst was malignant, and

the remaining five non-solid masses were benign. The most common margin descriptor for malignant masses was ill-defined (46%), and that for benign masses was circumscribed (59%). Most of the malignant masses had irregular shape (59%) and most of the benign masses had oval shape (70%). Most of the masses (76% of benign masses and 64% of malignant masses) were categorized as hypoechoic. Calcifications were seen in 2% of benign masses and 25% of malignant masses.

Table 1 shows the individual radiologist's A_z and $A_z^{(0.9)}$ values with and without CAD, and the two-tailed p-values for the change in both accuracy measures with CAD. The A_z values of the radiologists were in the range between 0.81 to 0.87 without CAD, and 0.86 to 0.93 with CAD. R4 had the largest change in A_z value when reading in the aided condition, with A_z values of 0.82 and 0.93 without and with CAD. The improvement in A_z was statistically significant for each individual radiologist.

The average ROC curves for the radiologists with and without CAD were derived from the average a and b parameters, which were defined as the means of the individual radiologist's a and b parameters for the fitted ROC curves. The average ROC curves are shown in Figure 4 along with the test ROC curve of the computer classifier, which had an A_z value of 0.92. Table 2 lists the average A_z and $A_z^{(0.9)}$ values, and the corresponding two-tailed p values estimated using the DBM method or the Student's paired t-test. The average A_z value improved significantly ($p<0.01$) from 0.84 to 0.90, and the average $A_z^{(0.9)}$ value improved significantly ($p=0.015$) from 0.30 to 0.47 with CAD. The improvement in the Az and $A_z^{(0.9)}$ values were statistically significant ($p<0.01$) even when R4, who showed the largest improvement with CAD, was excluded from the analysis.

The ROC curves of R1, R2 and R3 in our previous study (15) were compared to those in the current study without CAD both as a group (using the DBM method) and individually. The average A_z values for these three radiologists were 0.87 and 0.84 in the previous and current studies, respectively. The difference between the current and previous studies did not achieve statistical significance either as a group ($p=0.17$) or when each radiologist's ROC curves were analyzed separately ($p=0.80$, 0.13, and 0.09 for R1, R2, and R3, respectively.)

The sensitivity and specificity of each radiologist with and without CAD at an LM threshold of 2% are listed in Table 3. On the average, the radiologists' sensitivity increased from 96% to 98% with CAD, at the cost of a decrease in specificity from 22% to 19%. Three of the radiologists showed an increase in sensitivity while two maintained a sensitivity of 100%. The specificity of three radiologists decreased with CAD, while one radiologists' specificity increased and one did not show any change. The change in the sensitivity and specificity did not achieve statistical significance for any radiologist (range of p-values using McNemar's test: 0.157-1.00 for sensitivity, and 0.102-1.00 for specificity). Table 3 also shows the sensitivity and specificity for each radiologist if the LM threshold were to be adjusted to 7% when they read with CAD, for which the average sensitivity would remain at 96% (same as that without CAD) while the average specificity would increase to 46%. Under this condition, the improvement in specificity for 4 out of 5 radiologists was statistically significant ($p<0.002$, McNemar's test), while the change in sensitivity for each radiologist was insignificant.

With 102 cases and five radiologists, we had a total of 510 pairs of LM ratings with and without CAD. Figure 5 shows a histogram of the change in the radiologists'

LM ratings with CAD for these 510 readings. The radiologists did not change their LM rating substantially (i.e., within 5) with CAD in 64% (326/510) of the readings. For malignant masses, the ratings were substantially increased for 34% (95/280) and decreased for 7% (19/280) of the readings. For benign masses, the ratings were substantially increased for 14% (32/230) and decreased for 17% (38/230) of the readings.

Figure 6 shows the histogram of the mean change in the LM ratings for malignant and benign masses. To obtain the mean change for a mass, the changes with CAD from five radiologists were averaged. To statistically evaluate the change for malignant and benign masses, we performed one-sample t-tests on the mean changes. For benign masses, the decrease in the average LM rating was 0.77, which did not achieve statistical significance (two-tailed $p=0.51$). The increase in the average LM rating of malignant masses was 5.59, which was statistically significant (two-tailed $p<0.0001$).

The correlations between the radiologists' readings with and without CAD, and those between radiologists' readings with CAD and computer scores are shown in Table 4. The former of these correlations was higher than the latter for all five radiologists, and the difference between the two was statistically significant ($p<10^{-6}$) for four radiologists. This result indicates that when they read with CAD, radiologists had a higher agreement with their reading without CAD compared to their agreement with computer scores.

As described at the beginning of this section, 96 masses were categorized as solid by the majority rule. To investigate how the radiologists performed with and without CAD for solid masses, we applied ROC analysis to this subset by excluding cases that were categorized as cysts. The average A_z values without and with CAD for this subset

were 0.84 and 0.90, respectively, unchanged from the entire set of 102 masses. The improvements in A_z for the individual radiologists as well as for all radiologists as a group were statistically significant ($p<0.05$) for the subset of solid masses.

DISCUSSION

Our results indicate that the CAD algorithm used in this study was able to assist even expert breast imaging radiologists in characterizing masses on 3D US volumes. At our institution, all clinical breast US examinations are performed by breast imaging radiologists, not sonographers, and therefore the readers in our ROC study are particularly experienced in assessing whole volume images. Nevertheless, our CAD system could still improve their accuracy in terms of the A_z and $A_z^{(0.9)}$ values. The average A_z value improved significantly ($p=0.005$) from 0.84 to 0.90, and the average $A_z^{(0.9)}$ value improved significantly ($p=0.015$) from 0.30 to 0.47. The area under the ROC curve for the computer classifier ($A_z=0.92$) was higher than those of all radiologists without CAD in the study. With CAD, all radiologists showed a significant improvement in their A_z values, and one radiologist's A_z value surpassed that of the computer classifier.

During our observer experiment, 96/102 (94%) of the masses were assessed as solid according to the majority rule. When the analysis was limited to this subset of solid masses, the A_z values with and without CAD, and the significance of the improvement with CAD were essentially unchanged compared to the results with the entire data set of 102 cases. This indicates that CAD would be helpful even if we only considered the interpretation of the more difficult category of solid masses.

The effect of CAD was mixed when measured in terms of the radiologists' sensitivity and specificity values at the current threshold of biopsy recommendation (LM of 2%). With CAD, the average sensitivity of the five radiologists increased from 96% to 98%, while their average specificity for this data set decreased from 22% to 19%. Since all lesions except one in our data set underwent biopsy or fine needle aspiration after clinical imaging, the relatively low specificity of the radiologists with or without CAD is not unexpected. With the malignancy ratings recorded in the observer experiment, we can analyze whether these changes in the specificity and sensitivity reflect only a shift in decision threshold along the same ROC curve. If that was the case, CAD would not actually increase the diagnostic accuracy of radiologists but only change their decision threshold. The same effect would have been achieved by urging them to use a more lenient threshold for biopsy recommendation, without the need of CAD. As evidenced from the significant improvement in the ROC curves, our malignancy rating data strongly suggest that this is not the case. Although the changes in sensitivity or specificity did not reach statistical significance due to the relative small data set available in this study, these observations indicate a promising trend that may be achieved with CAD.

The ultimate clinical utility of a CAD system that results in an increased sensitivity at the cost of decreasing specificity depends on a cost/benefit analysis of the different correct and incorrect decisions. Alternatively, by appropriate training, it may be possible to translate the benefits with CAD into biopsy decisions that surpass unaided reading in terms of both sensitivity and specificity, or an improvement in specificity

without reducing sensitivity. For example, for our data set, if the threshold for biopsy with CAD could be changed to an LM rating of 7%, the average specificity with CAD would be improved to 46%, compared to 22% without CAD, while the average sensitivity would remain at 96% as noted above.

Since the "cost" of failing to biopsy a malignant lesion is much greater than that of a benign biopsy, it can logically be expected that radiologists may tend to use the CAD system to confirm and increase their LM estimate of malignant lesions while not easily reducing the LM estimate of low suspicion lesions. This will result in an overall increase in radiologists' LM ratings. Figure 5 suggests that this is indeed the case in our study. While the ratings for malignant masses demonstrated a stronger trend to increase than to decrease with CAD, the ratings for benign masses did not show a strong trend either way. It is also noted that the radiologists' ratings showed little or no change (less than 5%) for a large percentage (64%) of the masses. It therefore appears that radiologists tend to be very conservative in downgrading the LM of a lesion. As a result, the observed improvement in the radiologists' accuracy in this study was obtained mainly from an increase in the LM ratings of malignant masses. This led to an increase in sensitivity and a slight decrease in specificity. However, since the ROC curves of all radiologists did improve with CAD, there is a potential that the radiologists can adjust their decision thresholds along the higher ROC curves and thus increase the sensitivity as well as the specificity. Alternatively, it may be possible to convince them to reduce the LM ratings of masses that the CAD system rates as very low suspicion, thus improving the specificity. These improvements may be realized after radiologists accumulate experiences and increase their confidence with the use of CAD.

The assessment of mass characteristics during the observer study helped us better understand the properties of our data set. These assessments were not used in designing the computer classifier since they were obtained during the observer experiment after the classifier design was already completed. The assessments did not directly influence radiologists' LM ratings, but may have served to focus their attention on a systematic and thorough evaluation of the characteristics of the mass. It was reported that a systematic analysis of the characteristics of breast lesions guided by a checklist could improve radiologists' diagnostic accuracy (28). The list of mass descriptors collected in this study is similar to that in the ultrasound BI-RADS lexicon recently published by the American College of Radiology (20). However, since the BI-RADS lexicon for breast US had not been published at the time of the study, the descriptors are not exactly the same.

As explained in the Methods Section 28/130 cases (21%) were excluded from the study. The majority of the excluded cases were either simple cysts or cases which were deemed technically unsuccessful (e.g., those that contain partial mass volumes or motion artifacts). The number of technically unsuccessful cases was relatively large compared to that in clinical practice because the 3D US scanner is an experimental system. We expect that the computer classifier may not perform well for the scans that were technically unsuccessful because of the low quality of the data. By the same token, radiologists will refrain from making a clinical decision with images containing technical problems. We excluded these cases in the current study because it was not meaningful to train a computer classifier to deal with technical problems that will have been resolved if the 3D US scanner is to be implemented clinically in the future. We excluded the cases

with simple cysts because radiologists can already distinguish simple cysts from other lesions with very high accuracy using current US criteria. It is unlikely that they need the CAD system to aid them in this task. The inclusion of relatively easy cases may also optimistically bias the classifier.

A number of research groups have been developing CAD systems for breast mass characterization on US images in recent years. (12-15). Chen et al. (13) used morphological features extracted from hand-segmented mass boundaries on 2D US images to design a nearly setting-independent classifier. Using an automated segmentation method, Horsch et al. (14) obtained an A_z value of 0.87 in the task of differentiating all malignant and benign lesions (N=400) in their 2D US data set, and 0.82 in the task of differentiating the subset of malignant and benign solid lesions (N=276). Sahiner et al. (15) designed a classifier based on features extracted from 3D US images, and found that the accuracy of the designed classifier in estimating the likelihood of malignancy of masses was similar to that of experienced radiologists when their performances were compared for the same set of images. These previous studies, therefore, indicate that computer classifiers can perform well for characterizing masses on US images, although it is not possible to directly compare the performances of the classifiers because they were tested on different data sets. However, we are aware of very few studies that investigated the effect of CAD for US mass characterization on radiologists' accuracy. Recently, Horsch et al. (29) found that the accuracy of both expert mammographers and community radiologists improved significantly when they read 2D US images with CAD. Our study differs from that by Horsch et al. in that 3D

US images were used, but our results reinforce their finding that experienced radiologists can benefit from reading US images with CAD.

In order to provide a reference of the computer performance to the radiologists, Gaussian probability density distributions fitted to the computer scores for the malignant and benign classes were displayed for the radiologists during the study. The display of these distributions is one of the methods to show the users what the computer scores mean, and to provide the radiologists with some method for mentally calibrating these scores. Ideally, if a large data set is available, one can train the classifier on a first data set, obtain the distribution of malignant and benign scores on an independent second set, and perform the observer study on a third set that is independent from the set for which the distributions are obtained. Our study design assumes that the test performance of the designed classifier is generalizable so that the distributions on the second and third hypothetical sets above are identical. This is by far the most common approach for laboratory ROC studies of the effects of CAD (4, 30, 31) because of the limited data set available. Further work is warranted to investigate the validity of this assumption.

The radiologists were not informed of the prevalence of cancer in the data set. However, they would likely assume that the prevalence of the disease was higher than that in the diagnostic population in clinical practice. This is because most laboratory ROC studies are designed to have approximately equal number of positive and negative cases in order to increase the statistical power for the same total number of cases read (23) and our observers are familiar with ROC studies. Gur et al found that no significant effects could be measured for prevalence in the range of 2% to 28% in laboratory ROC experiments (32). It is not known if their findings could be extended to prevalence of

nearly 50%. On the other hand, since ROC studies are usually performed to measure the relative performances of two (or more) modalities instead of their absolute performances in the patient population at large, the prevalence effects should be comparable for both modalities and would be unlikely to change the relative performances, as assumed in most laboratory ROC studies.

When radiologists perform a laboratory experiment with CAD, there may be a concern that they may overly rely on the computer results, especially if they know that the computer classifier is very accurate. However, our observations indicated that the radiologists did not develop a tendency to follow the computer ratings in this study. First, as we discussed in the Results section, the radiologists did not change their LM rating substantially (i.e., greater than or equal to 5 on the 100 point scale) with CAD in 64% (326/510) of the readings. Second, the correlation analysis (Table 4) indicated that the LM ratings of a radiologist with and without CAD were highly correlated, whereas the correlation between the computer scores and the radiologists' LM ratings with CAD was significantly lower for four readers. Third, before all the readings were completed and the ROC analysis performed, radiologists did not know if their A_z was lower or higher than that of the computer. They also did not have feedback after reading a case regarding whether the computer's rating was more accurate than their rating. The radiologist thus had no way to know that they would improve by simply following the computer.

The US images used for analysis by the CAD system in this study constituted a volume that contained the biopsy-proven mass, acquired using an experimental system. The radiologists in our observer study were asked to characterize the masses based on

the same US volumes. In clinical practice, typically, these readers will interactively optimize the image quality by changing the probe angle, direction, and US scan settings for a given mass. The images interpreted in our observer performance study were therefore different from those our radiologists routinely interpret. The potentially less-than-optimal image quality may have had a negative impact on their reading accuracy. To our knowledge, all CAD systems developed so far for breast US operate on static images, and therefore do not take advantage of the interactive nature of US imaging. The use of 3D volumes for CAD design may reduce this disadvantage by providing a more complete description of the mass compared to a few 2D images that slice through the mass. Similarly, interpretation of 3D US volumes by a radiologist may offer advantages compared to interpretation of only a few hardcopy images acquired by a US technologist, although interactive acquisition by a radiologist may still be the best approach. Although current CAD systems have been designed for off-line processing of recorded US images to facilitate algorithm development in the laboratory, it is conceivable that the processing may be sped up to real time or within seconds of the US exam by firmware implementation in the future to make it compatible with clinical operations.

Our study had a number of limitations. As described in the Introduction, one of the purposes of our CAD system was to help radiologists reduce the benign biopsy rate without affecting the sensitivity of breast cancer detection. Our data set therefore consisted of only masses that were recommended for biopsy or fine needle aspiration. However, if such a system were used prospectively, it may affect the management of cases that the radiologist would normally recommend for a follow-up. It is therefore

important in the future to investigate the performance of the CAD system for masses that are not recommended for biopsy, and whose outcomes are known by follow-up. A second limitation is that all the cases in our data set were collected using the same US machine. Although we believe that our image processing methods will not depend strongly on small changes in image quality of the US images, the CAD system needs to be evaluated with images acquired using different US imaging systems to ensure its robustness against variations in image acquisition systems and parameters. A third limitation is that all the observers in our study were very experienced in breast imaging and US interpretation so that the effects of CAD on less experienced radiologists are still unknown. Although we believe that less experienced readers may benefit from CAD at least as much as the experienced radiologists, if not more, it will be important to investigate the effects of CAD on radiologists with mixed experiences.

Another limitation of our study is that the classifier in our CAD system was trained and tested using a leave-one-case-out method, and the segmentation method was optimized using a small subset of the data set. In the leave-one-case-out method for classifier design, the features are selected and the classifier is designed using N-1 cases, and the designed classifier is applied to the left-out case to determine the test score. Test scores for each case are obtained in round-robin order. Although this is known as a nearly unbiased classifier design method (19), the performance of our CAD system needs to be evaluated using independent test sets in order to assure the generalizability of our approach. Nevertheless, despite the need to confirm that the computer classifier results are generalizable, our observer study revealed the potential benefits that CAD may

provide to the radiologists for the characterization of masses, if a CAD system with the level of performance used in our study is available as a second opinion.

Finally, radiologists generally combine information from US with that from mammograms to reach a diagnostic decision while the current study only used the information from US images. The effects of CAD on a combined US and mammogram evaluation remain to be investigated. In addition to these limitations, retrospective ROC studies cannot emulate many factors that exist in clinical practice such as the psychological effects of the liability of missing a malignant case. The results observed in laboratory ROC studies thus may not be generalizable to clinical settings. However, ROC observer studies have been established as the one of the best available methods to-date to compare the relative performances of different imaging modalities or conditions. A laboratory ROC study is therefore an important first step to assess the effectiveness of CAD for assisting radiologists in making diagnostic decisions and may provide pilot data for the design of future clinical trials.

ACKNOWLEDGMENTS

This work was supported in part by U.S. Army Medical Research Materiel Command grant DAMD17-01-1-0328 and by USPHS grants CA095153 and CA091713. The authors are grateful to Charles E. Metz, Ph.D., for the LABMRMC program.

APPENDIX

Feature extraction

The feature vector for a given mass consisted of four width-to-height features, four posterior shadowing features, and 72 texture features.

The width-to-height features for a mass were the minimum, maximum, mean, and the standard deviation of the ratio of the width to the height of the segmented mass for each slice containing the mass. The width W and height H of the segmented mass in a slice were defined as the widest and the tallest cross-sections of the mass in that slice, respectively (Figure 2).

The posterior shadowing features for a mass were the minimum, maximum, mean, and the standard deviation of the feature extracted from each slice containing the mass. On a given slice, the posterior region of the mass was divided into n overlapping vertical strips as shown in Figure 2. The width of each strip was equal to $W/4$, and the height of the strip was equal to H. The strips were defined only posterior to the central $3W/4$ portion of the mass so that bilateral shadows that are sometimes associated with fibroadenomas could be avoided. Let P denote the mean grayscale value within the darkest posterior strip, and M denote the mean grayscale value within the segmented mass. The difference D between M and P defined how dark the US image is in the darkest posterior strip of the mass compared to the average within the mass. The posterior shadowing feature for the slice was defined as the normalized difference D/M.

The texture features were extracted from disc-shaped regions posterior and anterior to the mass. These equal-sized regions contained partly the interior portion of the mass and partly the mass margins. The total area of the anterior and posterior regions

was equal to the area of the segmented mass. An example of the anterior disc-shaped region is shown in Figure 2. On each slice containing the mass, spatial gray level dependence (SGLD) matrices, $S(d, \theta)$ were extracted. The $(i,j)^{th}$ element of $S(d, \theta)$ is the relative frequency with which two pixels, one with gray level i and the other with gray level j, separated by a pixel pair distance d in a direction θ, occur in the image. In this study, three pixel pair distances, d=2, 4, and 6, and two pixel pair angles, θ=0° and 90° were used. On each slice, we therefore extracted six SGLD matrices from the anterior and six SGLD matrices from the posterior disc-shaped regions. From each SGLD matrix, six texture features were extracted. These features were information measures of correlation 1 and 2, entropy, difference entropy, sum entropy, and energy. The mathematical definitions of these features can be found in the literature (33). The texture feature vector extracted from a slice was therefore 72-dimensional. These vectors were averaged over all slices containing a mass to obtain the texture feature vector for the mass.

REFERENCES

1. Kopans DB. The positive predictive value of mammography. AJR 1992; 158:521-526.
2. Adler DD, Helvie MA. Mammographic biopsy recommendations. Curr. Op. Radiol. 1992; 4:123-129.
3. Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. AJR 1995; 165:1373-1377.
4. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Gopal SS. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study. Radiology 1999; 212:817-827.
5. Huo ZM, Giger ML, Vyborny CJ, Metz CE. Breast cancer: Effectiveness of computer-aided diagnosis - Observer study with independent database of mammograms. Radiology 2002; 224:560-568.
6. Hadjiiski LM, Chan HP, Sahiner B, Helvie MA, Roubidoux M, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. Improvement of Radiologists' Characterization of Malignant and Benign Breast Masses in Serial Mammograms by Computer-Aided Diagnosis: An ROC Study. Radiology 2004; 233:255-265.
7. Jackson VP. The role of US in breast imaging. Radiology 1990; 177:305-311.

8. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: Use of sonography to distinguish between malignant and benign lesions. Radiology 1995; 196:123-134.

9. Skaane P, Engedal K. Analysis of sonographic features in differentiation of fibroadenoma and invasive ductal carcinoma. AJR 1998; 170:109-114.

10. Taylor KJW, Merritt C, Piccoli C, Schmidt R, Rouse G, Fornage B, Rubin E, Georgian-Smith D, Winsberg F, Goldberg B, Mendelson E. Ultrasound as a complement to mammography and breast examination to characterize breast masses. Ultrasound Med. Biol. 2002; 28:19-26.

11. Garra BS, Krasner BH, Horri SC, Ascher S, Mun SK, Zeman RK. Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis. Ultrasonic Imaging 1993; 15:267-285.

12. Chen DR, Chang RF, Huang YL. Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. Radiology 1999; 213:407-412.

13. Chen CM, Chou YH, Han KC, Hung GS, Tiu CM, Chiou HJ, Chiou SY. Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks. Radiology 2003; 226:504-514.

14. Horsch K, Giger ML, Venta LA, Vyborny CJ. Computerized diagnosis of breast lesions on ultrasound. Med. Phys. 2002; 29:157-164.

15. Sahiner B, Chan HP, Roubidoux MA, Helvie MA, Hadjiiski LM, Ramachandran A, LeCarpentier GL, Nees A, Paramagul C, Blane CE. Computerized characterization of breast masses on 3-D ultrasound volumes. Med. Phys. 2004; 31:744-754.

16. Bhatti PT, LeCarpentier GL, Roubidoux MA, Fowlkes JB, Helvie MA, Carson PL. Discrimination of sonographically detected breast masses using frequency shift color Doppler imaging in combination with age and gray scale criteria. Journal of Ultrasound in Medicine 2001; 20:343-350.

17. Carson PL, Fowlkes JB, Roubidoux MA, Moskalik AP, A. G, Normolle D, LeCarpentier GL, Nattakom S, Helvie MA, Rubin JM. 3-D color Doppler image quantification of breast masses. Ultrasound Med. Biol. 1998; 24:945-952.

18. Draper NR. Applied regression analysis. New York: Wiley, 1998.

19. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York: Springer-Verlag, 2001.

20. American College of Radiology Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). 4th ed. Reston, VA: American College of Radiology, 2003.

21. Sickles EA. Nonpalpable, circumscribed, noncalcified solid breast masses: likelihood of malignancy based on lesion size and age of patient. Radiology 1994; 192:439-442.

22. Beiden SV, Wagner RF, Doi K, Nishikawa RM, Freedman M, Lo S-C, Xu X-W. Independent versus Sequential Reading in ROC Studies of Computer-assist Modalities: Analysis of Component of Variance. Acad. Radiol. 2002; 9:1036-1043.

23. Metz CE. ROC methodology in radiologic imaging. Invest. Radiol. 1986; 21:720-733.

24. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 1984; 4:137-150.

25. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745-750.

26. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: Generalization to the population of readers and cases with the jackknife method. Invest. Radiol. 1992; 27:723-731.

27. Cohen J, Cohen P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillside, NJ: Lawrence Erlbaum, 1983.

28. Getty DJ, Pickett RM, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988; 23:240-252.

29. Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad. Radiol. 2004; 11:272-280.

30. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad. Radiol. 1999; 6:22-33.

31. Li F, Aoyama M, Shiraishi J, Abe H, Li Q, Suzuki K, Engelmann R, Sone S, MacMahon H, Doi aK. Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy. AJR Am J Roentgenol 2004; 183:1209-1215.

32. Gur D, Rockette HE, Armfield DR, Blachar A, Bogan JK, Brancatell G, Britton CA, Brown ML, Davis PL, Ferris JV, Fuhrman CR, Golla SK, Katyal S, Lacomis JM, McCook BM, Thaete FL, Warfel TE. Prevalence effect in a laboratory environment. Radiology 2003; 228:10-14.

33. Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans. Sys. Man. and Cybern. 1973; SMC-3:610-621.

TABLES

Table 1.

The area A_z under ROC curve, and the partial area index $A_z^{(0.9)}$ above a sensitivity of 0.9, for the characterization of the masses in the data set without and with CAD by the 5 radiologists.

Rad.	A_z			$A_z^{(0.9)}$		
No	No CAD	With CAD	p value	No CAD	With CAD	p value
1	0.83±0.04	0.89±0.03	0.0008	0.25±0.10	0.35±0.14	0.17
2	0.81±0.04	0.86±0.04	0.0005	0.14±0.08	0.23±0.12	0.13
3	0.87±0.03	0.91±0.03	0.0486	0.39±0.12	0.53±0.12	0.0747
4	0.82±0.04	0.93±0.02	0.0004	0.39±0.10	0.68±0.09	0.0008
5	0.83±0.04	0.90±0.03	0.0007	0.29±0.10	0.42±0.12	0.0323

Note —The A_z and $A_z^{(0.9)}$ values are the mean ± SD. The statistical significance for each radiologist was estimated as described in the literature (24, 25).

Table 2.

The average A_z and $A_z^{(0.9)}$ values without and with CAD for the five radiologists, obtained by using the average a and b parameters from the fitted ROC curves.

Accuracy measure	No CAD	With CAD	p value (DBM)	p value (paired t-test)
A_z	0.84	0.90	0.006	0.005
$A_z^{(0.9)}$	0.30	0.47	---	0.015

Note — The significance of the change in the A_z value with CAD for the group of five radiologists was estimated using both the DBM method and the Student's two-tailed paired t-test. The significance of the change in the $A_z^{(0.9)}$ value was estimated using the Student's two-tailed paired t-test.

Table 3.

The sensitivity and specificity for each radiologist at decision thresholds of 2% and 7% likelihood of malignancy.

Rad.	Sensitivity			Specificity		
No.	No CAD*	With CAD*	With CAD**	No CAD*	With CAD*	With CAD**
1	56 (100)	56 (100)	56 (100)	4 (9)	5 (11)	15 (33)
2	51 (91)	53 (95)	49 (88)	12 (26)	11 (24)	28 (61)
3	52 (93)	54 (96)	53 (95)	24 (52)	22 (48)	29 (63)
4	55 (98)	56 (100)	56 (100)	9 (20)	5 (11)	23 (50)
5	56 (100)	56 (100)	56 (100)	1 (2)	1 (2)	11 (24)
Avg.	54 (96)	55 (98)	54 (96)	10 (22)	9 (19)	21 (46)

Note — In each entry, the first number denotes the number of correctly classified lesions, and the number in parentheses denotes the percentage (i.e., sensitivity for the first three columns. and the specificity for the last three columns). The total numbers of malignant and benign lesions are 56 and 46, respectively.

* The columns entitled "No CAD*" and "With CAD*" show the sensitivity and specificity at the decision threshold of 2% likelihood of malignancy, without and with CAD, respectively.

** The columns entitled "With CAD**" show the sensitivity and specificity with CAD at a hypothetical decision threshold of 7% likelihood of malignancy, for which the average sensitivity would be the same as that without CAD (96%), but the average specificity would increase to 46%.

Table 4.

Correlations between the radiologists' LM ratings with and without CAD, and those between radiologists' LM ratings with CAD and computer scores. The statistical significance in the difference between the two correlation coefficients of each radiologist was estimated using Cohen and Cohen's method (27).

Rad. No	Correlation between LM ratings with CAD and		p-value
	LM ratings w/o CAD	Computer scores	
1	0.94	0.70	$<10^{-6}$
2	0.96	0.61	$<10^{-6}$
3	0.96	0.72	$<10^{-6}$
4	0.86	0.83	0.26
5	0.94	0.70	$<10^{-6}$

CAPTIONS FOR ILLUSTRATIONS

Figure 1: Five slices containing a malignant mass and the result of computer segmentation.

Figure 2: For feature extraction, the width *W* and height *H* of the mass on a slice were defined as the widest and the tallest cross-sections of the mass in that slice, respectively. The mean gray level values within the overlapping posterior strips *R(i)* and the segmented mass were used to define the posterior shadowing features. The disc-shaped regions for texture feature extraction followed the shape of the mass and contained partly the segmented mass and partly its margins. An example of the anterior disc-shaped region is shown as the gray area above the segmented mass.

Figure 3: The graphical user interface. The biopsy-proven lesion was marked by an arrow, which could be switched off when the radiologist assessed the mass. The interface allowed the users to navigate through the volume, and to adjust the contrast and brightness. The radiologists first provided their assessment for the mass in six categories, which were 1) overall US impression; 2) shape; 3) margins; 4) echogenicity; 5) through transmission; and 6) other features. They then provided a likelihood of malignancy rating without CAD. Finally, the computer's malignancy score for the mass was displayed and the radiologists had an option to revise their rating after taking into consideration the computer's opinion.

Figure 4: The average ROC curves of the radiologists with and without CAD, and the ROC curve of the computer classifier. The average ROC curves were constructed by using the mean a and b values of the individual observers' ROC curves.

Figure 5: The histogram of the change in radiologists' ratings with CAD. For the majority of the masses (59% of malignant masses and 70% of benign masses) the change was in the range of -4 to 4. When the change in the scores with CAD was greater than or equal to the range of -5 to 5, the change was called substantial. For malignant masses, the ratings were substantially increased for an average of 34% (95/280) and decreased for 7% (19/280) of the readings. For benign masses, the ratings were substantially increased for 14% (32/230) and decreased for 17% (38/230) of the readings.

Figure 6: The histogram of the mean change in the LM ratings of radiologists with CAD. The mean change for a mass was computed by averaging the changes in the LM ratings for that mass over the five radiologists who participated in the study. For benign masses, the overall average LM rating decrease was 0.77, which did not achieve statistical significance ($p=0.51$). For malignant masses the overall average LM rating increase was 5.59, which was statistically significant ($p<0.0001$).

ILLUSTRATIONS

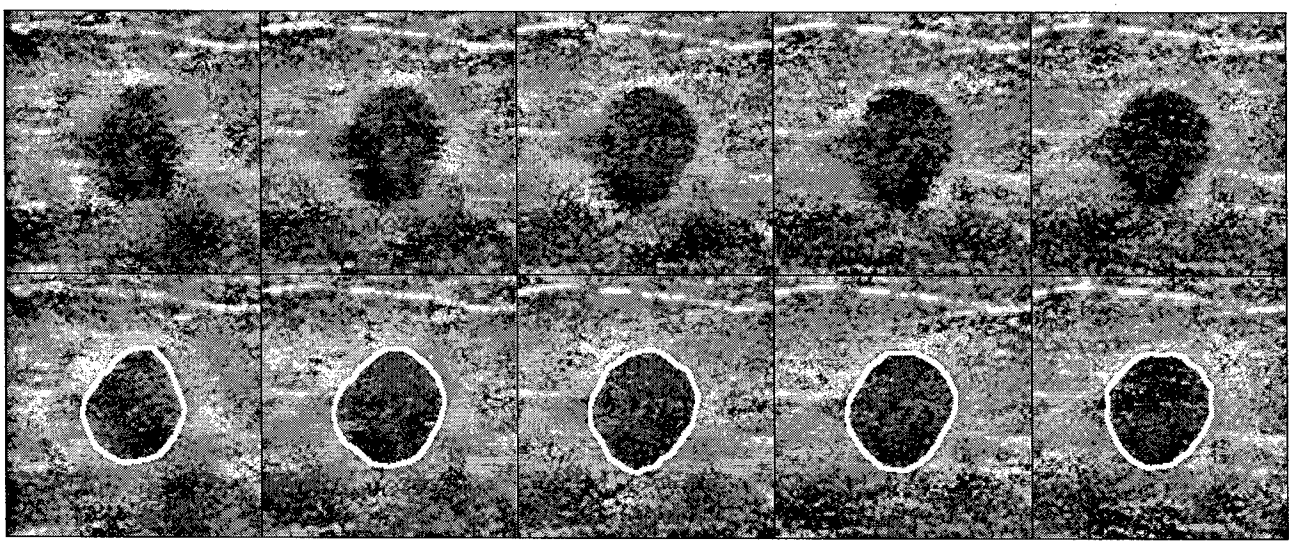


Figure 1: Five slices containing a malignant mass and the result of computer segmentation.

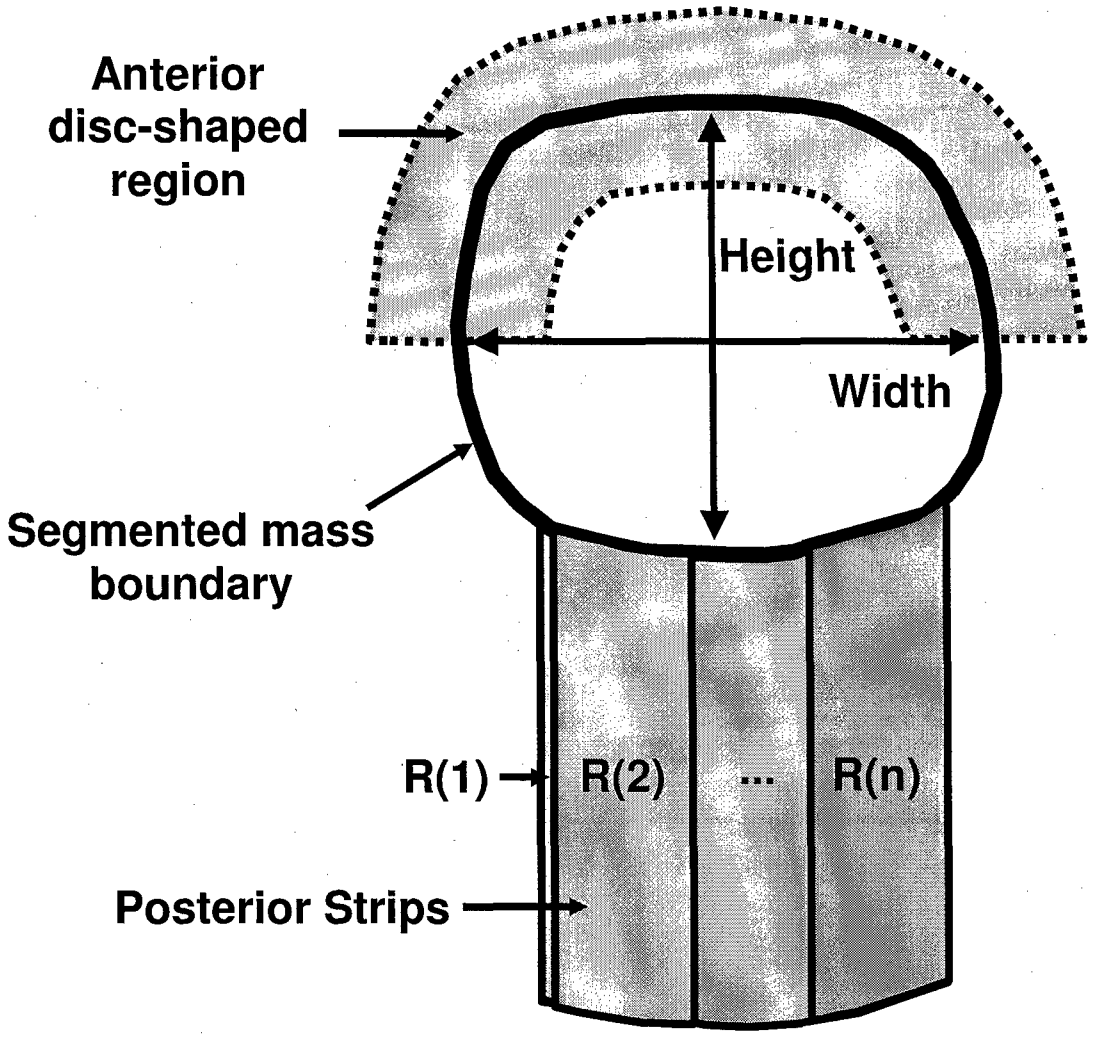


Figure 2: For feature extraction, the width *W* and height *H* of the mass on a slice were defined as the widest and the tallest cross-sections of the mass in that slice, respectively. The mean gray level values within the overlapping posterior strips *R(i)* and the segmented mass were used to define the posterior shadowing features. The disc-shaped regions for texture feature extraction followed the shape of the mass and contained partly the segmented mass and partly its margins. An example of the anterior disc-shaped region is shown as the gray area above the segmented mass.

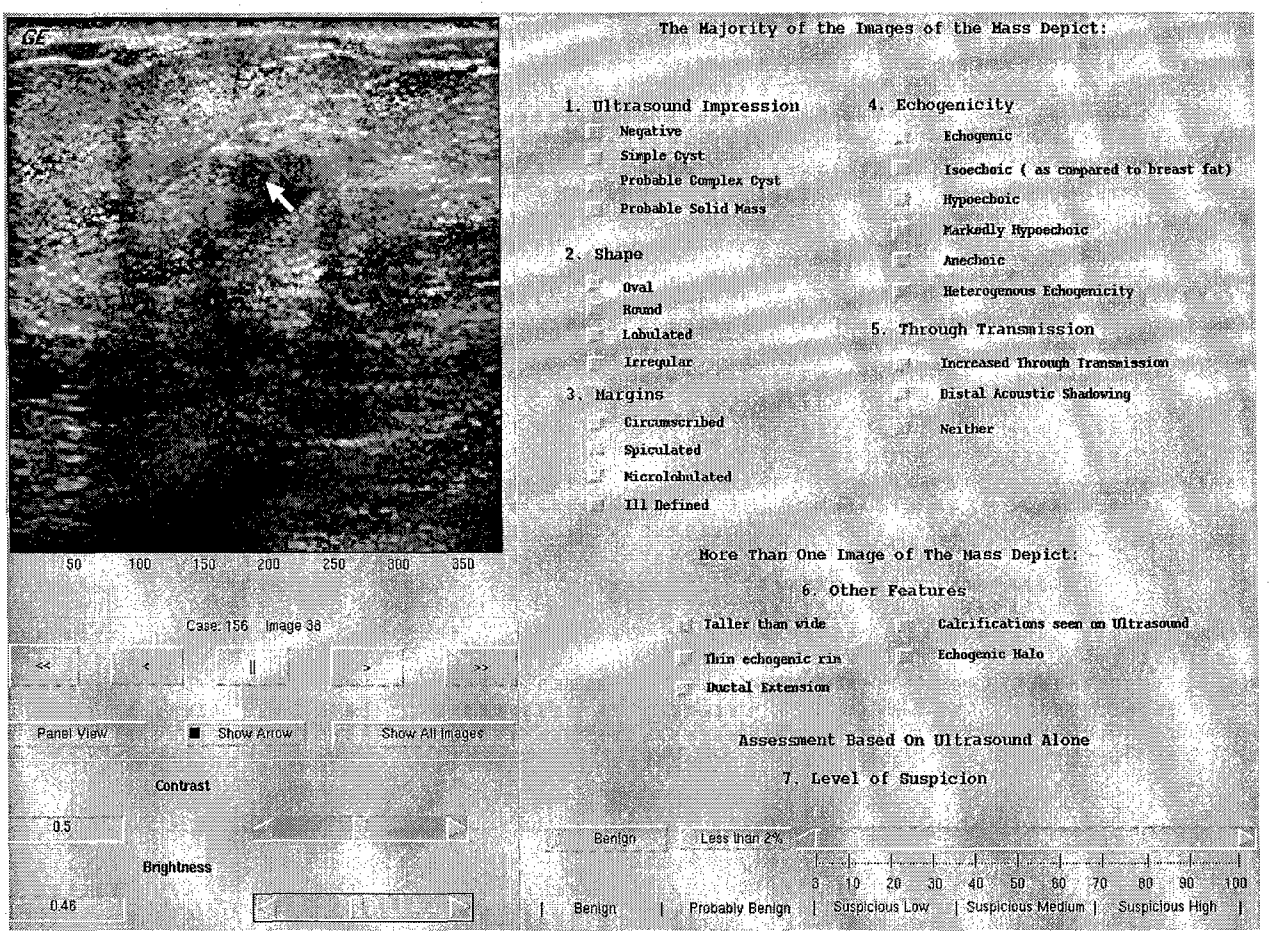


Figure 3: The graphical user interface. The biopsy-proven lesion was marked by an arrow, which could be switched off when the radiologist assessed the mass. The interface allowed the users to navigate through the volume, and to adjust the contrast and brightness. The radiologists first provided their assessment for the mass in six categories, which were 1) overall US impression; 2) shape; 3) margins; 4) echogenicity; 5) through transmission; and 6) other features. They then provided a likelihood of malignancy rating without CAD. Finally, the computer's malignancy score for the mass was displayed and the radiologists had an option to revise their rating after taking into consideration the computer's opinion.

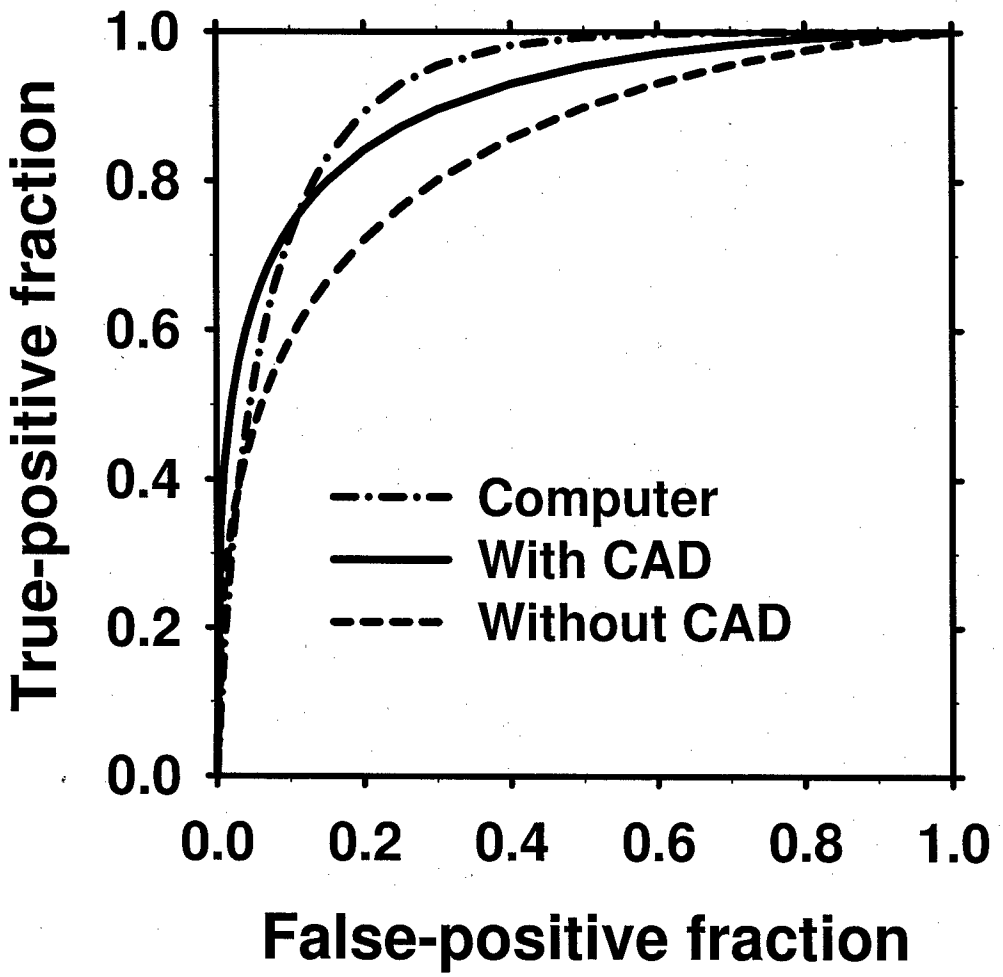


Figure 4: The average ROC curves of the radiologists with and without CAD, and the ROC curve of the computer classifier. The average ROC curves were constructed by using the mean a and b values of the individual observers' ROC curves.

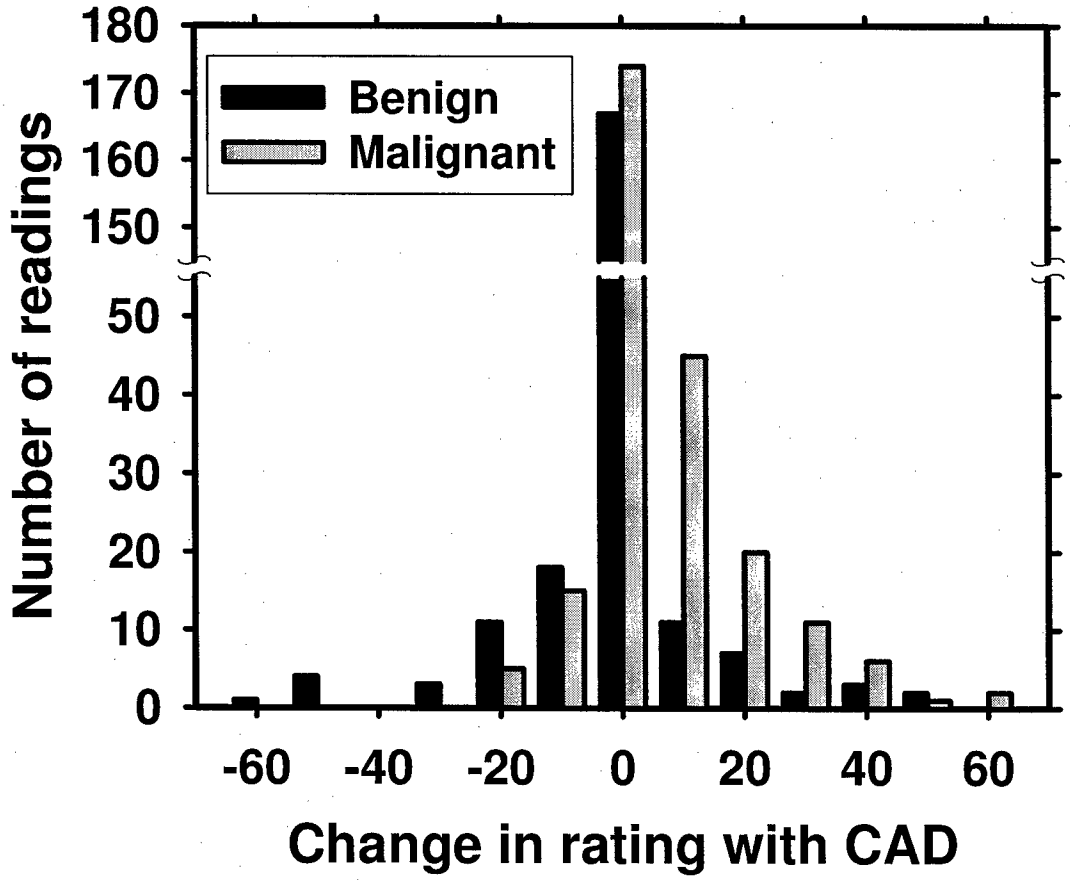


Figure 5: The histogram of the change in radiologists' ratings with CAD. For the majority of the masses (59% of malignant masses and 70% of benign masses) the change was in the range of -4 to 4. When the change in the scores with CAD was greater than or equal to the range of -5 to 5, the change was called substantial. For malignant masses, the ratings were substantially increased for an average of 34% (95/280) and decreased for 7% (19/280) of the readings. For benign masses, the ratings were substantially increased for 14% (32/230) and decreased for 17% (38/230) of the readings.

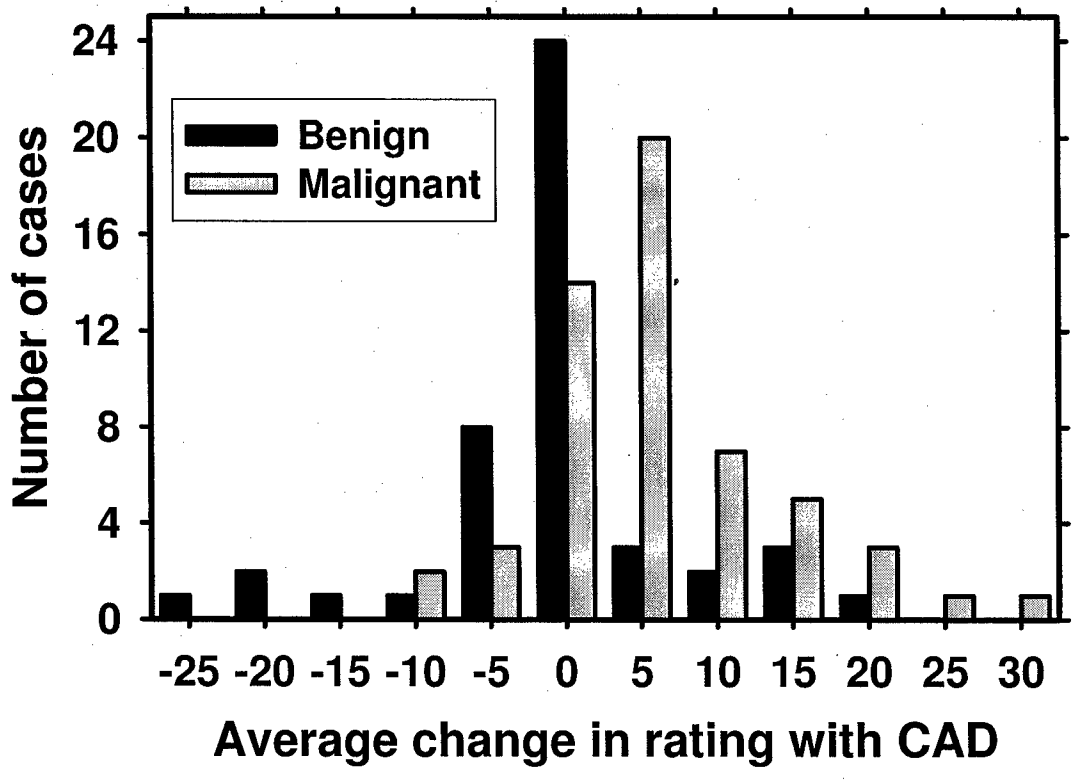


Figure 6: The histogram of the mean change in the LM ratings of radiologists with CAD. The mean change for a mass was computed by averaging the changes in the LM ratings for that mass over the five radiologists who participated in the study. For benign masses, the overall decrease in the average LM rating was 0.77, which did not achieve statistical significance ($p=0.51$). For malignant masses the overall increase in the average LM rating was 5.59, which was statistically significant ($p<0.0001$).

Appendix 2

Abstract ID: 4412549
Submission Type: Scientific Papers

Berkman Sahiner
University of Michigan Medical Center
Phone: 734-647-7429
Fax: 734-615-5513
E-Mail: berki@umich.edu

THE EFFECT OF A MULTI-MODALITY COMPUTER CLASSIFIER ON RADIOLOGISTS' ACCURACY IN CHARACTERIZING BREAST MASSES USING MAMMOGRAMS AND VOLUMETRIC ULTRASOUND IMAGES: AN ROC STUDY

B Sahiner (P); H Chan; L M Hadjiiski; M A Roubidoux; C P Paramagul; M A Helvie ; et al.

PURPOSE
Computer-aided diagnosis (CAD) methods have previously been developed to assist radiologists in characterizing breast masses on mammograms and ultrasound (US) images. In this study, we developed a classifier that merged information from both modalities, and assessed its effect on radiologists' accuracy.

METHOD AND MATERIALS
The data set consisted of images from 67 patients containing biopsy-proven solid masses (32 benign and 35 malignant). An experienced radiologist identified the region of interest (ROI) containing the lesion on both modalities. The 3D US volumetric data were collected as cine-clips when the transducer was translated across the lesion. US and mammographic features were automatically extracted based on the margin, spiculation, shadowing, and shape characteristics of the masses. The features were combined into a malignancy score using a computer classifier designed with a leave-one-case-out method. Five MQSA radiologists participated in the ROC study. First, the radiologist read the mammogram ROIs, and provided a BIRADS score and a malignancy rating. Second, the US images were displayed along with the mammogram ROIs, the radiologist provided a second malignancy rating, and recommended: (i) 1-year follow-up; (ii) short-term follow-up; or (iii) biopsy. Third, the computer score was displayed, and the radiologist provided a third malignancy rating and revised the recommended action. The classification accuracy was quantified using the area under ROC curve, Az.

RESULTS
The computer classifier achieved a test Az value of 0.91. When reading mammograms alone, the radiologists had an average Az of 0.88 (range: 0.82-0.93). When the mammograms were supplemented by US images, the average Az increased to 0.92 (range:0.86-0.96). With CAD, the average Az increased significantly ($p=0.03$) to 0.95 (range:0.90-0.98). The average sensitivity for biopsy recommendation also improved from 0.96 to 0.98, and average specificity improved from 0.37 to 0.39.

CONCLUSIONS
The radiologists were more accurate in characterizing masses when both mammograms and volumetric US images were available. A well-trained computer algorithm can improve radiologists' accuracy even in this multi-modality reading condition.

Appendix 3

COMPUTER-AIDED MULTI-MODALITY BREAST MASS CHARACTERIZATION*

Berkman Sahiner, Heang-Ping Chan, Lubomir M. Hadjiiski, Marilyn A. Roubidoux, Chintana Paramagul, Mark A. Helvie

*Presented at the fourth Era of Hope meeting for the Department of Defense (DOD) Breast Cancer Research Program (BCRP), June 8-11, 2005, Philadelphia, Pennsylvania.

Abstract

Ultrasound (US) and mammography are two commonly used modalities for characterization of breast masses. We are developing techniques for computerized characterization of masses on these two modalities, and for the fusion of the computer-extracted information. Our goal is to provide a consistent and reliable computer-based second opinion to radiologists that may improve their accuracy in characterizing breast masses.

We have developed automated segmentation algorithms based on 2D and 3D active contour models for the segmentation of the masses on mammograms and 3D US images. US features that may be useful for characterizing masses as malignant or benign were extracted based on the margin, shadowing, and shape characteristics of the mass. Mammographic features were extracted based on texture, morphological, and spiculation characteristics.

We have investigated the accuracy of a classifier based on computer-extracted mammographic features alone, US features alone, and the combined feature space. The accuracy of the designed classifier was evaluated using receiver operating characteristic (ROC) methodology. The area A_z under the test ROC curve for the computer classifier using the US images alone, mammograms alone, and the combined feature space were 0.88±0.04, 0.86±0.05, and 0.91±0.03, respectively.

We have also investigated the effect of the designed multi-modality classifier on radiologists' accuracy in characterizing masses. Ten MQSA radiologists participated in an ROC study. First, the radiologist read the region of interest (ROI) on the mammograms that contained the mass, and provided a BI-RADS score and a malignancy rating. Second, the US images were displayed in addition to the mammogram ROIs, and the radiologist provided a second malignancy rating, and recommended either follow-up or biopsy. Third, the CAD results were displayed, and the radiologist provided a third malignancy rating and revised the recommended action. With CAD, the average A_z increased significantly ($p=0.05$) from 0.93 to 0.95. The average sensitivity for biopsy recommendation improved from 0.96 to 0.99, and average specificity improved from 0.27 to 0.29. Alternatively, if a biopsy recommendation threshold could be chosen such that the sensitivity were maintained at 98%, the specificity would increase to 39% with CAD.

Our results indicate that the designed multi-modality classifier significantly increases the accuracy of radiologists' assessment of masses. Our system therefore has a potential to be a valuable clinical tool for reducing the biopsy of benign lesions without a trade-off in sensitivity.

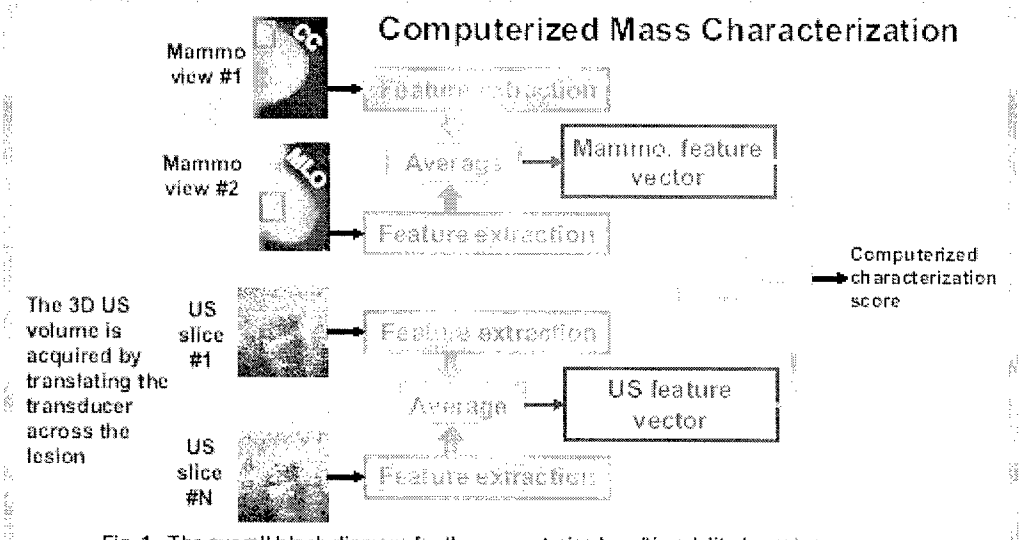


Fig. 1. The overall block diagram for the computerized multimodality breast mass classifier that uses multi-view mammograms and 3D ultrasound volumes.

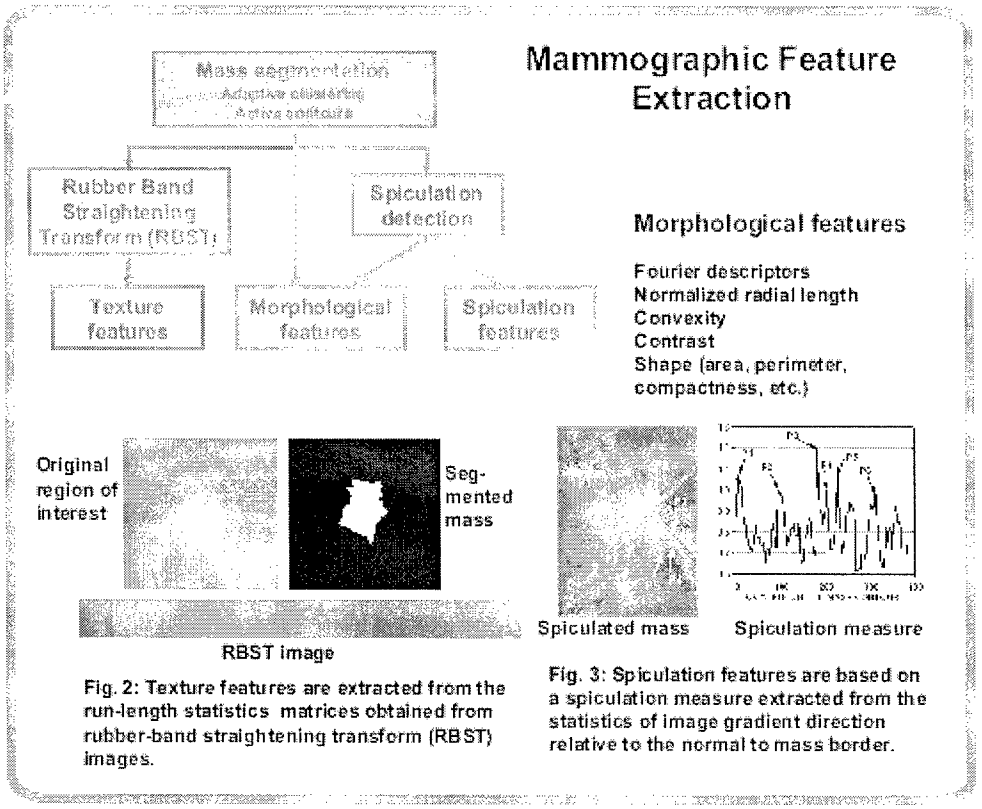


Fig. 2: Texture features are extracted from the run-length statistics matrices obtained from rubber-band straightening transform (RBST) images.

Fig. 3: Spiculation features are based on a spiculation measure extracted from the statistics of image gradient direction relative to the normal to mass border.

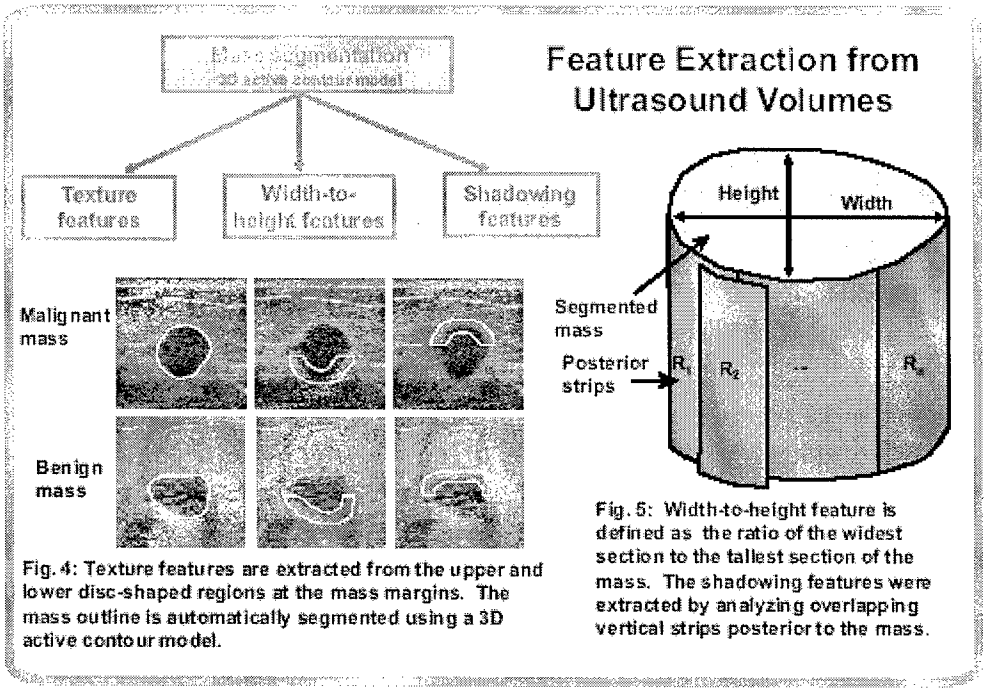


Fig. 4: Texture features are extracted from the upper and lower disc-shaped regions at the mass margins. The mass outline is automatically segmented using a 3D active contour model.

Fig. 5: Width-to-height feature is defined as the ratio of the widest section to the tallest section of the mass. The shadowing features were extracted by analyzing overlapping vertical strips posterior to the mass.

Computerized Characterization Results

Our data set consisted of US volumes and mammograms from 67 patients who had a mammographically visible solid mass deemed suspicious or highly suggestive of malignancy. All patients underwent biopsy or fine needle aspiration. Thirty two of the masses were benign and 35 were malignant. The total number of mammographic views was 163, with each case containing between one and three views (CC, MLO, or LAT). The biopsied mass on the mammograms and the US volumes was identified by an MQSA (Mammography Quality Standards Act) qualified radiologist using clinical images and case reports to confirm that the identified region contained the biopsied mass. The mammographic and ultrasound features were combined into a malignancy score using a stepwise linear discriminant analysis classifier (Fig. 1), trained and tested using a leave-one-case-out method. The malignancy scores were analyzed using Receiver Operating Characteristic (ROC) methodology. The area A_z under the ROC curve was used as the measure of accuracy.

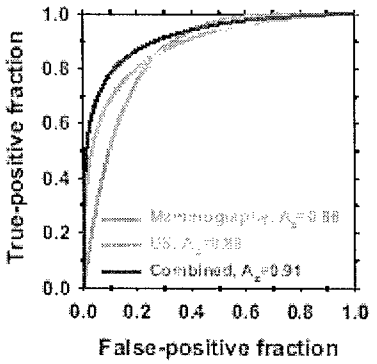


Fig. 6: The ROC curves for the computer classifier using mammograms alone, US volume alone, and the combination of both modalities.

Modality	A_z
Mammography	0.86±0.05
US	0.86±0.04
Combined	0.91±0.03

Observer Performance Study

We conducted an observer performance study to investigate the effect of the multimodality computer classifier on radiologists' accuracy in characterizing breast masses. Ten MQSA radiologists evaluated the images in randomized reading order. First, the radiologist read the mammogram ROIs, and provided a BI-RADS score and a likelihood of malignancy rating. Second, the US images were displayed along with the mammogram ROIs, the radiologist provided a second malignancy rating, and recommended: (i) 1-year follow-up; (ii) short-term follow-up; or (iii) biopsy. Third, the computer score was displayed, and the radiologist provided a third malignancy rating and revised the recommended action. The radiologist ratings were analyzed using Dorfman-Berbaum-Metz multiple reader multiple case (MRMC) ROC analysis.

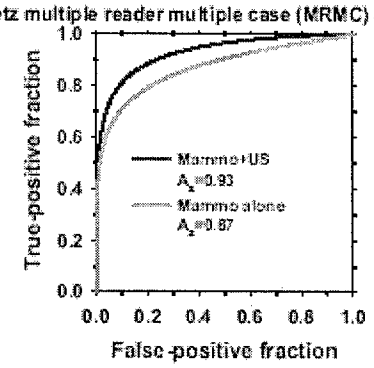


Fig. 7: The average ROC curves of radiologists interpreting mammograms alone, and mammograms supplemented by US volumes. The improvement with US was statistically significant (p=0.03).

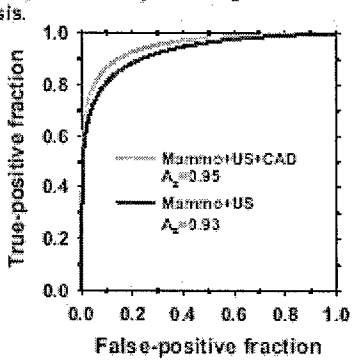
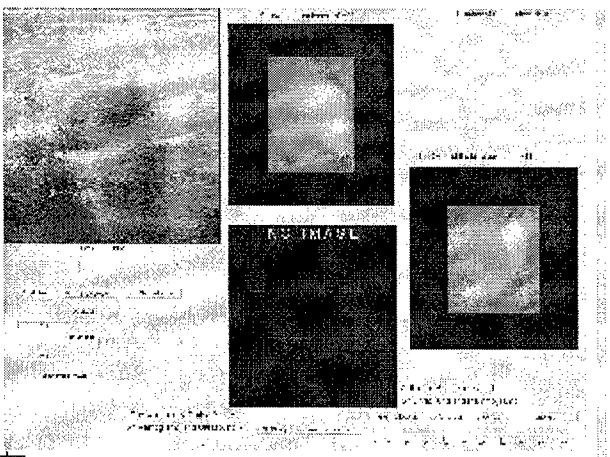


Fig. 8: The average ROC curves of radiologists interpreting mammograms and US volumes without and with CAD. The improvement with CAD was statistically significant (p=0.05).

Observer Performance Study

Fig. 9: The graphical user interface designed for the observer study experiment. The user can scroll through the US volume in cine mode, and change settings such as contrast and brightness for each image. After the observer provides a likelihood of malignancy rating for the case, the CAD score is displayed and the observer has an option to change his/her rating.



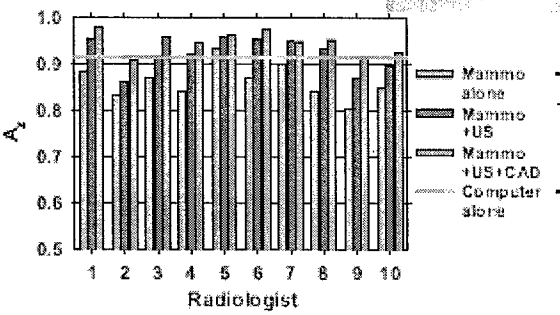


Fig. 10: The A_z values of each radiologist under the three reading conditions. The accuracy of each radiologist improved with US. Each radiologist except one improved further when they read with CAD.

Reading condition	Sensitivity	Specificity
Mammo alone	0.94	0.33
Mammo+US	0.98	0.27
Mammo+US+CAD	0.98	0.39

Fig. 11: The sensitivity and specificity for each reading condition based on the BI-RADS rating or the action category. Alternatively, if a threshold could be selected so that the sensitivity with CAD were maintained at 98%, the specificity with CAD would increase to 39%.

Conclusion

We have investigated a multi-modality computer classifier for characterization of breast masses on mammograms and 3D US images. The combination of the two modalities increased the accuracy of the computer classifier compared to that using each modality alone. Similarly, radiologists were significantly more accurate when the mammograms were supplemented with US images for interpretation. Our study demonstrated that even when expert radiologists use both modalities for interpretation, CAD may still have an important role to play. The radiologists' accuracy were significantly improved when they used CAD under this reading condition. When they read with CAD, the accuracy of nine out of ten radiologists improved compared to that without CAD, and the A_z value of eight radiologists was higher than that of the computer algorithm. CAD improved radiologists' average sensitivity from 98% to 99%, and average specificity from 27% to 29%. Alternatively, if a threshold could be selected so that the sensitivity with CAD were maintained at 98% (same as reading without CAD), the specificity would increase to 39% with CAD. Our results therefore demonstrate that CAD has the potential to assist radiologists in reducing the number of benign biopsies without decreasing the sensitivity of breast cancer detection.

Acknowledgment

This work was supported in part by U.S. Army Medical Research Materiel Command grant DAMD17-01-1-0328.